

LOAN DOCUMENT

PHOTOGRAPH THIS SHEET

①

INVENTORY

LEVEL

DTIC ACCESSION NUMBER

AFOSR-TR-92-0704

DOCUMENT IDENTIFICATION

25 Jun 92

IN PRESSION STATION. A

Revised for release;

Distribution: Unlimited

DISTRIBUTION STATEMENT

ACCESSION FOR

NTIS GRA&I

DTIC TRAC

UNANNOUNCED

JUSTIFICATION

BY

DISTRIBUTION/

AVAILABILITY CODES

DISTRIBUTION

AVAILABILITY AND/OR SPECIAL

A-1

DISTRIBUTION STAMP

DTIC QUALITY INSPECTED 5

DATE RECEIVED IN DTIC

DATE ACCESSIONED

DATE RETURNED

390743 1113 pg

92-20724



REGISTERED OR CERTIFIED NUMBER

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-FDAC

H
A
N
D
L
E

W
I
T
H

C
A
R
E

AD-A254 457



**UNITED STATES AIR FORCE
1989 RESEARCH INITIATION PROGRAM**

**Conducted by
UNIVERSAL ENERGY SYSTEMS, INC.**

**under
USAF Contract Number F49620-88-C-0053**

RESEARCH REPORTS

VOLUME I OF IV

**Submitted to
Air Force Office of Scientific Research
Bolling Air Force Base**

Washington, DC

**By
Universal Energy Systems, Inc.**

June 1992

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE 25 Jun 92		3. REPORT TYPE AND DATES COVERED ANNUAL 1 Jan 90 to 31 Dec 90	
4. TITLE AND SUBTITLE US Air Force 1989 Research Initiation Program Conducted by Universal Energy Systems, Inc, VOL # /				5. FUNDING NUMBERS F49620-88-C-0053 2305/D5	
6. AUTHOR(S) Rod Darrah					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Universal Energy Systems, Inc Dayton OH				8. PERFORMING ORGANIZATION REPORT NUMBER AFOSR-TR-92-004	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/NI Bolling AFB DC				10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES					
12a. DISTRIBUTION / AVAILABILITY STATEMENT (U)				12b. DISTRIBUTION CODE (U)	
13. ABSTRACT (Maximum 200 words) <p>This program is for follow-on research efforts for the participants in the Summer Faculty Research Program. Funding is provided to establish RIP awards to about half the number of participants in the SFRP. Participants in the 1989 SFRP competed for funding under the 1989 RIP. Evaluation of the proposals were made by the contractor. Evaluation criteria consisted of: 1. Technical excellence of the proposal 2. Continuation of the SFRP effort 3. Cost sharing by the university. The list of proposals selected for award was forwarded to AFOSR for approval of funding and for research efforts to be completed by 31 December 1990. The following summarizes the events for the evaluation of proposals and award of funding under the RIP. A. RIP proposals were submitted to the contractor by 1 November 1990. The proposals were limited to \$20,000 plus cost sharing by the universities. The universities were encouraged to cost share, since this is an effort to establish a long term effort between the Air Force and the university. B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories. C. Subcontracts were negotiated with the Universities. There were a total of 122 RIP awards made under the 1989 program.</p>					
14. SUBJECT TERMS				15. NUMBER OF PAGES	
				16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT (U)		18. SECURITY CLASSIFICATION OF THIS PAGE (U)		19. SECURITY CLASSIFICATION OF ABSTRACT (U)	
20. LIMITATION OF ABSTRACT (U)					

**UNITED STATES AIR FORCE
1989 RESEARCH INITIATION PROGRAM**

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

under

USAF Contract Number F49620-88-C-0053

RESEARCH REPORTS

VOLUME I OF IV

Submitted to

Air Force Office of Scientific Research

Bolling Air Force Base

Washington, DC

By

Universal Energy Systems, Inc.

June 1992

TABLE OF CONTENTS

<u>SECTION</u>	<u>PAGE</u>
INTRODUCTION	ii
STATISTICS	iii
PARTICIPANT LABORATORY ASSIGNMENT	vii
RESEARCH REPORTS	xv

INTRODUCTION

Research Initiation Program - 1989

AFOSR has provided funding for follow-on research efforts for the participants in the Summer Faculty Research Program. Initially, this program was conducted by AFOSR and popularly known as the Mini-Grant Program. Since 1983 the program has been conducted by the Summer Faculty Research Program (SFRP) contractor and is now called the Research Initiation Program (RIP). Funding is provided to establish RIP awards to about half the number of participants in the SFRP.

Participants in the 1989 SFRP competed for funding under the 1989 RIP. Participants submitted cost and technical proposals to the contractor by 1 November 1989, following their participation in the 1989 SFRP.

Evaluation of these proposals were made by the contractor. Evaluation criteria consisted of:

1. Technical excellence of the proposal
2. Continuation of the SFRP effort
3. Cost sharing by the university

The list of proposals selected for award was forwarded to AFOSR for approval of funding. Those approved by AFOSR were funded for research efforts to be completed by 31 December 1990.

The following summarizes the events for the evaluation of proposals and award of funding under the RIP.

- A. RIP proposals were submitted to the contractor by 1 November 1989. The proposals were limited to \$20,000 plus cost sharing by the universities. The universities were encouraged to cost share, since this is an effort to establish a long term effort between the Air Force and the university.
- B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories.
- C. Subcontracts were negotiated with the universities. The period of performance of the subcontract was between October 1989 and December 1990.

Copies of the final reports are presented in Volumes I through IV of the 1989 Research Initiation Program Report. There were a total of 122 RIP awards made under the 1989 program.

STATISTICS

PROGRAM STATISTICS

Total SFRP Participants	168
Total RIP Proposals submitted by SFRP	132
Total RIP Proposals submitted by GSRP	2
Total RIP Proposals submitted	134
Total RIP's funded to SFRP	94
Total RIP's funded to GSRP	2
Total RIP's funded	96
Total RIP Proposals submitted by HBCU's	9
Total RIP Proposals funded to HBCU's	5

LABORATORY PARTICIPATION

<u>Laboratory</u>	<u>Participants</u>	<u>Submitted</u>	<u>Funded</u>
AAMRL	12	10	6
WRDC/APL	10	8	6
ATL	9	9 (1 GSRP)	9 (1 GSRP)
AEDC	10	8	8
WRDC/AL	7	5	4
ESMC	0	0	0
ESD	3	2	1
ESC	11	8	7
WRDC/FDL	9	7	5
FJSRL	7	5	4
AFGL	12	10	6
HRL	12	10 (1 GSRP)	8 (1 GSRP)
WRDC/ML	9	7	5
OEHL	4	1	1
AL	12	10	6
RADC	15	11	8
SAM	17	16	9
WL	8	7	3
WHMC	1	0	0
Total	168	134	96

LIST OF PARTICIPATING UNIVERSITIES

Alabama, University of	- 1	New York, State University of	- 2
Alfred University	- 1	North Carolina State University	- 1
Arkansas-Pine Bluff, Univ. of	- 1	Northern Arizona University	- 1
Auburn University	- 1	Northern Illinois University	- 1
Bethel College	- 1	Northwestern University	- 1
Boston College	- 1	Notre Dame, University of	- 1
Brescia College	- 1	Ohio State University	- 2
California Polytechnic	- 1	Oklahoma, University of	- 3
California State University	- 2	Old Dominion University	- 1
Cincinnati, University of	- 2	Pennsylvania State University	- 1
Denver, University of	- 1	Pittsburgh, University of	- 1
Eastern Kentucky University	- 1	Rhode Island, University of	- 1
Florida Atlantic University	- 1	San Diego State University	- 1
Florida Institute	- 1	San Jose State University	- 1
Florida, University of	- 4	Savannah State College	- 1
Hamilton College	- 1	Scranton, University of	- 1
Harvard University	- 1	Southern Oregon State College	- 2
Illinois Institute of Technology	- 1	Southwest Texas State University	- 1
Illinois-Rockford, University of	- 1	Tennessee State University	- 1
Illinois State University	- 1	Tennessee Technological Univ.	- 1
Indiana-Purdue, University of	- 1	Texas A&M University	- 6
Kansas State University	- 2	Texas Southern University	- 1
Lawrence Technological University	- 1	Texas-San Antonio, University of	- 3
Long Island University	- 1	Transylvania University	- 1
Lowell, University of	- 1	Trinity University	- 1
Massachusetts, University of	- 2	US Naval Academy	- 1
Michigan, University of	- 1	Utah State University	- 1
Minnesota-Duluth, University of	- 2	Utica College	- 1
Mississippi State University	- 2	Vanderbilt University	- 1
Missouri-Rolla, University of	- 1	Washington State University	- 1
Murray State University	- 1	West Virginia University	- 2
Nebraska-Lincoln, University of	- 2	Wisconsin-Platteville, Univ. of	- 1
New Hampshire, University of	- 1	Worcester Polytechnic Institute	- 1
New York Institute of Technology	- 1	Wright State University	- 6
		Total	- 94

PARTICIPANTS LABORATORY ASSIGNMENT

AERO PROPULSION AND POWER DIRECTORATE

(Wright-Patterson Air Force Base)

Dr. Jerry Clark
Wright State University
Specialty: Physics

Dr. Frank Gerner
University of Cincinnati
Specialty: Mechanical Engineering

Dr. Thomas Lalk
Texas A&M University
Specialty: Mechanical Engineering

Dr. Baruch Lieber
State University of New York
Specialty: Aerospace Engineering

Dr. William Schulz
Eastern Kentucky University
Specialty: Analytical Chemistry
760-7MG-079 and 210-10MG-095

Dr. Richard Tankin
Northwestern University
Specialty: Mechanical Engineering

ARMAMENT DIRECTORATE

(Eglin Air Force Base)

Dr. Peter Armendarez
Brescia College
Specialty: Physical Chemistry

Dr. Joseph Brown
Mississippi State University
Specialty: Mechanical Engineering

Dr. Roger Bunting
Illinois State University
Specialty: Inorganic Chemistry

Dr. Satish Chandra
Kansas State University
Specialty: Electrical Engineering

Dr. David Cicci
Auburn University
Specialty: Aerospace Engineering

Mr. William Newbold (GSRP)
University of Florida
Specialty: Aerospace Engineering

Dr. Boghos Sivazlian
University of Florida
Specialty: Operations Research

Dr. Steven Trogon
University of Minnesota-Duluth
Specialty: Mechanics

Mr. Asad Yousuf
Savannah State College
Specialty: Electrical Engineering

ARMSTRONG LABORATORY

(Brooks Air Force Base)

Dr. Robert Blystone
Trinity University
Specialty: Zoology

Dr. Carolyn Caudle-Alexander
Tennessee State University
Specialty: Microbiology

Dr. James Chambers
University of Texas - San Antonio
Specialty: Biochemistry

Dr. Mark Cornwall
Northern Arizona University
Specialty: Human Performance

Dr. Vito DelVecchio
University of Scranton
Specialty: Biochemical Engineering

Dr. Gwendolyn Howze
Texas Southern University
Specialty: Molecular Biology

Dr. Harold Longbotham
University of Texas-San Antonio
Specialty: Electrical Engineering

Dr. Ralph Peters (1987)
Wichita State University
Specialty: Zoology

Dr. Raymond Quock
Univ. of Illinois at Rockford
Specialty: Pharmacology

Dr. Ram Tripathi
University of Texas-San Antonio
Specialty: Statistics

ARNOLD ENGINEERING DEVELOPMENT CENTER

(Arnold Air Force Base)

Dr. Brian Beecken
Bethel College
Specialty: Physics

Dr. Stephen Cobb
Murray State University
Specialty: Physics

Dr. John Francis
University of Oklahoma
Specialty: Mechanical Engineering

Dr. Orlando Hankins
University of North Carolina State
Specialty: Nuclear Engineering

Dr. Lang-Wah Lee
University of Wisconsin-Platteville
Specialty: Mechanical Engineering

Dr. Chun Fu Su
Mississippi State University
Specialty: Physics

Dr. Richard Tipping
University of Alabama
Specialty: Physics

Dr. D. Wilkes
Vanderbilt University
Specialty: Electrical Engineering

AVIONICS DIRECTORATE

(Wright-Patterson Air Force Base)

Dr. David Choate
Transylvania University
Specialty: Mathematics

Dr. R. H. Cofer
Florida Institute
Specialty: Electrical Engineering

Dr. Dar-Biau Liu
California State University
Specialty: Applied Mathematics

Dr. Robert Shock
Wright State University
Specialty: Mathematics

CREW SYSTEMS DIRECTORATE

(Wright-Patterson Air Force Base)

Dr. Thomas Lockwood
Wright State University
Specialty: Toxicology

Dr. Ethel Matin
Long Island University
Specialty: Experimental Psychology

Dr. Randy Pollack
Wright State University
Specialty: Anthropology

Dr. Donald Robertson (1987)
Indiana University of Pennsylvania
Specialty: Psychology

Dr. Michael Stanisic
University of Notre Dame
Specialty: Robotics

Dr. Chi-Ming Tang
State University of New York
Specialty: Mathematics

Dr. Ebo Tei
University of Arkansas-Pine Bluff
Specialty: Psychology

ENGINEERING AND SERVICES CENTER

(Tyndall Air Force Base)

Dr. William Bannister
University of Lowell
Specialty: Organic Chemistry

Dr. Emerson Besch
University of Florida
Specialty: Animal Physiology

Dr. Avery Demond
University of Massachusetts
Specialty: Civil Engineering

Dr. Kirk Hatfield
University of Florida
Specialty: Civil Engineering

Dr. Kim Hayes
University of Michigan
Specialty: Environmental Engineering

Dr. Deborah Ross
University of Indiana-Purdue
Specialty: Microbiology

Dr. Dennis Truax (1987)
Mississippi State University
Specialty: Civil Engineering

Dr. George Veyera
University of Rhode Island
Specialty: Civil Engineering

ELECTRONIC SYSTEMS DIVISION

(Hanscom Air Force Base)

Dr. Stephen Kolitz (1986)
University of Massachusetts
Specialty: Operations Research

Dr. Sundaram Natarajan
Tennessee Technical University
Specialty: Electrical Engineering

FLIGHT DYNAMICS DIRECTORATE

(Wright-Patterson Air Force Base)

Dr. Kenneth Cornelius
Wright State University
Specialty: Fluid Mechanics

Dr. William Wolfe
Ohio State University
Specialty: Engineering

Dr. Arnold Polak
University of Cincinnati
Specialty: Aerospace Engineering

Dr. Lawrence Zavodney
Ohio State University
Specialty: Mechanical Engineering

Dr. Nisar Shaikh
University of Nebraska-Lincoln
Specialty: Applied Mathematics

FRANK J. SEILER RESEARCH LABORATORY

(United States Air Force Academy)

Dr. Robert Granger
US Naval Academy
Specialty: Mechanical Engineering

Dr. Timothy Troutt
Washington State University
Specialty: Mechanical Engineering

Dr. Clay Sharts
San Diego State University
Specialty: Chemistry

Dr. Hung Vu
California State University
Specialty: Applied Mechanics

GEOPHYSICS DIRECTORATE

(Hanscom Air Force Base)

Dr. Phanindramohan Das
Texas A&M University
Specialty: Geophysical Science

Dr. Thomas Miller
University of Oklahoma
Specialty: Physics

Dr. Alan Kafka
Boston College
Specialty: Geophysics

Dr. Henry Nebel
Alfred University
Specialty: Physics

Dr. Charles Lishawa
Utica College
Specialty: Physical Chemistry

Dr. Craig Rasmussen
Utah State University
Specialty: Physics

HUMAN RESOURCES DIRECTORATE

(Brooks, Williams and Wright-Patterson Air Force Base)

Dr. Kevin Bennett
Wright State University
Specialty: Applied Psychology

Mr. John Williamson (GSRP)
Texas A&M University
Specialty: Psychology

Dr. Deborah Mitta
Texas A&M University
Specialty: Industrial Engineering

Dr. Michael Wolfe
West Virginia University
Specialty: Management Science

Dr. William Smith
University of Pittsburgh
Specialty: Linguistics

Dr. Yehoshua Zeevi
Harvard University
Specialty: Electrical Engineering

Dr. Stanley Stephenson
Southwest Texas State University
Specialty: Psychology

Dr. Robert Zerwekh
Northern Illinois University
Specialty: Philosophy

MATERIALS DIRECTORATE

(Wright-Patterson Air Force Base)

Dr. Donald Chung
San Jose State University
Specialty: Material Science

Dr. Kenneth Currie
Kansas State University
Specialty: Industrial Engineering

Dr. Michael Resch
University of Nebraska-Lincoln
Specialty: Materials Science

Dr. James Sherwood
University of New Hampshire
Specialty: Aerospace Mechanics
210-9MG-088 and 210-10MG-098

Dr. Michael Sydor
University of Minnesota-Duluth
Specialty: Physics

OCCUPATIONAL AND ENVIRONMENTAL HEALTH DIRECTORATE

(Brooks Air Force Base)

Dr. Stewart Maurer
New York Institute of Technology
Specialty: Electrical Engineering

ROCKET PROPULSION DIRECTORATE

(Edwards Air Force Base)

Dr. Lynn Kirms
Southern Oregon State College
Specialty: Organic Chemistry

Dr. Mark Kirms
Southern Oregon State College
Specialty: Organic Chemistry

Dr. Faysal Kolkailah
California Polytechnic
Specialty: Mechanical Engineering

Dr. Vittal Rao
University of Missouri-Rolla
Specialty: Control Systems

Dr. Larry Swanson
University of Denver
Specialty: Mechanical Engineering

Dr. Roger Thompson
Pennsylvania State University
Specialty: Engineering Mechanics

ROME LABORATORIES

(Griffiss Air Force Base)

Dr. Charles Alajajian
West Virginia University
Specialty: Electrical Engineering

Dr. Ian Grosse
University of Massachusetts
Specialty: Mechanical Engineering

Dr. Henry Helmken
Florida Atlantic University
Specialty: Physics

Dr. Michael Klein
Worcester Poly Institute
Specialty: Physics

Dr. William Kuriger
University of Oklahoma
Specialty: Electrical Engineering

Dr. Khaja Subhani
Lawrence Tech. University
Specialty: Electrical Engineering

Dr. David Sumberg (1987)
Rochester Institute of Tech.
Specialty: Physics

Dr. Donald Ucci
Illinois Institute of Technology
Specialty: Electrical Engineering

Dr. Kenneth Walter (1988)
Prairie View A&M University
Specialty: Chemical Engineering

Dr. James Wolper
Hamilton College
Specialty: Mathematics

WEAPONS DIRECTORATE

(Kirtland Air Force Base)

Dr. Harry Hogan
Texas A&M University
Specialty: Mechanical Engineering

Dr. Arkady Kheyfets (1988)
North Carolina State University
Specialty: Mathematical Physics

Dr. Duc Nguyen
Old Dominion University
Specialty: Civil Engineering

Dr. Duane Sanders
Texas A&M University
Specialty: Civil Engineering

RESEARCH REPORTS

MINI-GRANT RESEARCH REPORTS

<u>Technical Report Number</u>	<u>Title and Mini-Grant Number</u>	<u>Professor</u>
Volume I		
Rome Laboratories		
1	Optimal Design of Finite Wordlength FIR Digital Filters for an Analog Transversal Filter with Tap Weight Circuitry Defects Using Adaptive Modeling 210-10MG-123	Dr. Charles Alajajian
2	Automatic Adaptive Remeshing for Finite Element Reliability Assessment of Electronic Devices 210-10MG-129	Dr. Ian Grosse
3	Ionospherically-Induced Phase Distortion Across Wide-Aperture HF Phased Arrays 210-10MG-047	Dr. Henry Helmken
4	A Study of Interacting Tunneling Units with Possible Application to High Temperature Superconductors 210-10MG-057	Dr. Michael Klein
5	Reduced Bandwidth Binary Phase-Only Filters 210-10MG-052	Dr. William Kuriger
6	Computer Modeling of GaAs/AlGaAs MQW Devices for Optical Properties 210-10MG-107	Dr. Khaja Subhani
7	Fiber Optic Distribution System for Phased Array Antennas 760-7MG-113	Dr. David Sumberg (1987)
8	Continuation Study of a Communications Receiver for Spread Spectrum Signals 210-10MG-067	Dr. Donald Ucci
9	Development of a System to Deposit Thin Films of Titanium Carbide Using Atomic Layer Epitaxy 219-9MG-113	Dr. Kenneth Walter (1988)

- | | | |
|----|---|------------------|
| 10 | Neural Networks for Invariant Pattern Recognition
210-10MG-061 | Dr. James Wolper |
|----|---|------------------|

Arnold Engineering Development Center

- | | | |
|----|---|---------------------|
| 11 | The Performance of IR Detectors Illuminated
by Monochromatic Radiation
210-10MG-029 | Dr. Brian Beecken |
| 12 | Sodium Fluorescence Studies for Application to
RDV of Hypersonic Flows
210-10MG-076 | Dr. Stephen Cobb |
| 13 | Report Not Publishable At This Time
210-10MG-086 | Dr. John Francis |
| 14 | NOT PUBLISHABLE AT THIS TIME
210-10MG-134 | Dr. Orlando Hankins |
| 15 | An Experimental Approach for the Design of a
Mixer for an Arc Heater
210-10MG-027 | Dr. Lang-Wah Lee |
| 16 | No Report Submitted (1986)
760-6MG-099 | Dr. Arthur Mason |
| 17 | Laser-Induced Fluorescence of Nitric Oxide
210-10MG-054 | Dr. Chun Fu Su |
| 18 | Spectroscopic Monitoring of Exhaust Gases
210-10MG-099 | Dr. Richard Tipping |
| 19 | Transient Analysis of Parallel Distributed
Structurally Adaptive Signal Processing Systems
210-10MG-084 | Dr. D. Wilkes |

Electronic Systems Division

- | | | |
|----|---|------------------------------|
| 20 | Reliability in Satellite Communication Networks
760-6MG-094 | Dr. Stephen Kolitz
(1986) |
| 21 | Comparison of Testability Analysis Tools for USAF
210-10MG-065 | Dr. Sundaram Natarajan |

Engineering and Services Center

- | | | |
|----|--|----------------------------|
| 22 | Anomalous Effects of Water in Fire Fighting:
Facilitation of JP Fires by Azeotropic
Distillation Effects
210-10MG-115 | Dr. William Bannister |
| 23 | Effect of Simulated Jet Aircraft Noise on
Domestic Goats
210-10MG-119 | Dr. Emerson Besch |
| 24 | Migration of Organic Liquid Contaminants Using
Measured and Estimated Transport Properties
210-10MG-025 | Dr. Avery Demond |
| 25 | Laboratory Investigations of Subsurface
Contaminant Sorption Systems
210-10MG-064 | Dr. Kirk Hatfield |
| 26 | Effects of Surfactants on Partitioning of
Hazardous Organic Components of JP-4 Onto
Low Organic Carbon Soils
210-10MG-125 | Dr. Kim Hayes |
| 27 | Biodegradation of Hydrocarbon Components of
Jet Fuel JP-4
210-10MG-018 | Dr. Deborah Ross |
| 28 | 760-7MG-079; See 210-10MG-095
Report # 71
(Aero Propulsion and Power Directorate) | Dr. William Schulz |
| 29 | Pretreatment of Wastewaters Generated by
Firefighter Training Facilities
760-7MG-105 | Dr. Dennis Truax
(1987) |
| 30 | Stress Transmission and Microstructure in
Compacted Moist Sand
210-10MG-019 | Dr. George Veyera |

Frank J. Seiler Research Laboratory

- | | | |
|----|---|---------------------|
| 31 | No Report Submitted (1985)
760-0MG-008 | Dr. Hermann Donnert |
|----|---|---------------------|

- | | | |
|----|--|--------------------|
| 32 | Reference AIAA 91-0745; Flow Induced Vibrations of Thin Leading Edges; U.S. Naval Academy
210-10MG-011 | Dr. Robert Granger |
| 33 | No Report Submitted (1985)
760-0MG-107 | Dr. Ronald Sega |
| 34 | Use of Nitronium Triflate for Nitration of Nitrogen Heterocycles
210-10MG-072 | Dr. Clay Sharts |
| 35 | No Report Submitted (1985)
760-0MG-053 | Dr. Walter Trafton |
| 36 | Active Control of Dynamic Stall Phenomena
210-10MG-049 | Dr. Timothy Troutt |
| 37 | Modeling and Control of a Fundamental Structure-Control System: A Cantilever Beam and a Structure-Borne Reaction-Mass Actuator
210-10MG-021 | Dr. Hung Vu |

Volume II

Phillips Laboratory

Geophysics Directorate

- | | | |
|----|---|------------------------|
| 38 | Cumulus Parameterization in Numerical Prediction Models: A New Parcel-Dynamical Approach
210-10MG-087 | Dr. Phanindramohan Das |
| 39 | R _g as a Depth Discriminant for Earthquakes and Explosions in New England and Eastern Kazakhstan
210-10MG-082 | Dr. Alan Kafka |
| 40 | Time-of-Flight Simulations of Collisions of H ₂ ¹⁸ O ⁺ with D ₂ O
210-10MG-117 | Dr. Charles Lishawa |
| 41 | Electron Attachment to Transition-Metal Acids
210-10MG-113 | Dr. Thomas Miller |

- | | | |
|----|---|---------------------|
| 42 | CO2 (4.3 μ m) Vibrational Temperatures and Limb Radiances in the Mesosphere and Lower Thermosphere: Sunlit Conditions and Terminator Conditions
210-10MG-055 | Dr. Henry Nebel |
| 43 | Development and Application of a Dynamo Model of Electric Fields in the Middle-and Low-Latitude Ionosphere
210-10MG-060 | Dr. Craig Rasmussen |

Rocket Propulsion Directorate

- | | | |
|----|--|----------------------|
| 44 | Synthesis of Tetranitrohomocubane
210-10MG-091 | Dr. Lynn Kirms |
| 45 | Synthesis of Poly(Imide Siloxane) Copolymers and Graft Copolymers
210-10MG-090 | Dr. Mark Kirms |
| 46 | Finite Element Analysis for Composite Structures
210-10MG-127 | Dr. Faysal Kolkailah |
| 47 | Robust Control of Large Flexible Structures Using Reduced Order Models
210-10MG-043 | Dr. Vittal Rao |
| 48 | Theoretical Study of Capillary Pumping in Heat Pipes
210-10MG-026 | Dr. Larry Swanson |
| 49 | Multi-Body Dynamics Experiment Design
210-10MG-121 | Dr. Roger Thompson |

Advanced Weapons Survivability Directorate,
Lasers and Imaging Directorate, and
Space and Missile Technology Directorate

- | | | |
|----|---|-----------------------|
| 50 | No Report Submitted (1988)
210-9MG-119 | Dr. Lane Clark |
| 51 | No Report Submitted (1986)
760-6MG-054 | Dr. Fabian Hadipriono |

- | | | |
|----|--|-------------------------------|
| 52 | Improved Modeling of the Response of Pressurized Composite Cylinders to Laser Damage
210-10MG-008 | Dr. Harry Hogan |
| 53 | Relativistic Effects in Global Positioning
210-9MG-114 | Dr. Arkady Kheyfets
(1988) |
| 54 | No Report Submitted (1987)
760-7MG-047 | Dr. Barry McConnell |
| 55 | Parallel and Vector Processing for Nonlinear Finite Element Analysis
210-10MG-051 | Dr. Duc Nguyen |
| 56 | Resonant Scattering of Elastic Waves by a Random Distribution of Spherical Inclusions in a Granular Medium
210-10MG-085 | Dr. Duane Sanders |

Volume III

Wright Laboratory

Armament Directorate

- | | | |
|----|---|---------------------------|
| 57 | Reactive Aluminum "Burst"
210-10MG-106 | Dr. Peter Armendarez |
| 58 | Damage of Aircraft Runways by Aerial Bombs
210-10MG-104 | Dr. Joseph Brown |
| 59 | Ionic Polymer Membranes for Capacitor Electrolytes
210-10MG-096 | Dr. Roger Bunting |
| 60 | Multisensor Seeker Feasibility Study for Medium
Range Air-to-Air Missiles
210-10MG-074 | Dr. Satish Chandra |
| 61 | Sequential Ridge-Type Estimation Methods
210-10MG-044 | Dr. David Cicci |
| 62 | Numerical Simulation of Transonic Flex-Fin
Projectile Aerodynamics
210-10MG-005 | Mr. William Newbold |
| 63 | Effectiveness Models for Smart Submunitions
Systems
210-10MG-002 | Dr. Boghos Sivazlian |
| 64 | Detonation Modeling of Explosives Using the
Hull Hydrodynamics Computer Code
210-10MG-010 | Dr. Steven Trogon |
| 65 | Stress Analysis of a Penetrator using Finite
Element Method
210-9MG-015 | Dr. Wafa Yazigi
(1988) |
| 66 | Knowledge-Based Target Detection for the
RSPL/IPL Laboratories
210-10MG-017 | Mr. Asad Yousuf |

Aero Propulsion and Power Directorate

- | | | |
|----|---|-----------------|
| 67 | Study of Electron Impact Infrared Excitation
Furtions of Xenon
210-10MG-100 | Dr. Jerry Clark |
|----|---|-----------------|

- | | | |
|----|---|--------------------|
| 68 | Micro Heat Pipes
210-10MG-066 | Dr. Frank Gerner |
| 69 | No Report Submitted
210-10MG-109 | Dr. Thomas Lalk |
| 70 | Analysis of the Flowfield in a Pipe with a Sudden
Expansion and with Different Coaxial Swirlers
210-10MG-001 | Dr. Baruch Lieber |
| 71 | Jet Fuel Additive Efficiency Analysis with a
Surrogate JP-8 Fuel
210-10MG-095 | Dr. William Schulz |
| 72 | Comparison Between Experiments and Predictions
Based on Maximum Entropy for Sprays from a
Pressure Atomizer
210-10MG-036 | Dr. Richard Tankin |

Avionics Directorate

- | | | |
|----|--|------------------|
| 73 | An Algorithm to Resolve Multiple Frequencies
210-10MG-031 | Dr. David Choate |
| 74 | Model Based Bayesian Target Recognition
210-10MG-022 | Dr. R. H. Cofer |
| 75 | Study of Sky Backgrounds and Subvisual
Cirrus
210-9MG-120 | Dr. Gerald Grams |
| 76 | Simulation of Dynamic Task Scheduling
Algorithms for ADA Distributed System
Evaluation Testbed (ADSET)
210-10MG-020 | Dr. Dar-Biau Liu |
| 77 | Towards a Course-Grained Test Suite for VHDL
Validation
210-10MG-012 | Dr. Robert Shock |

Flight Dynamics Directorate

- | | | |
|----|--|-----------------------|
| 78 | Experimental Study of Pneumatic Jet/Vortical
Interaction on a Chined Forebody Configuration
at High Angles of Attack
210-10MG-046 | Dr. Kenneth Cornelius |
|----|--|-----------------------|

- | | | |
|----|---|-----------------------|
| 79 | Numerical Study of Surface Roughness Effect on Hypersonic Flow Separation
210-10MG-056 | Dr. Arnold Polak |
| 80 | Ultrasonic Stress Measurements and Craze Studies for Transparent Plastic Enclosures of Fighter Aircraft
210-10MG-126 | Dr. Nisar Shaikh |
| 81 | 210-9MG-088, See 210-10MG-098
Report # 87
Materials Directorate | Dr. James Sherwood |
| 82 | Experimental Determination of Damage Initiation Resulting from Low Velocity Impact of Composites
210-10MG-094 | Dr. William Wolfe |
| 83 | The Response of Nonlinear Systems to Random Excitation
210-10MG-093 | Dr. Lawrence Zavodney |

Materials Directorate

- | | | |
|----|--|--------------------|
| 84 | The In-Situ Deposition of High Tc Superconducting Thin Film by Laser Ablation
210-10MG-116 | Dr. Donald Chung |
| 85 | Self-Improving Process Control for Molecular Beam Epitaxy of Ternary Alloy Materials on GaAs and InPh Substrates
210-10MG-030 | Dr. Kenneth Currie |
| 86 | Detection of Fatigue Crack Initiation Using Surface Acoustic Waves
210-10MG-120 | Dr. Michael Resch |
| 87 | Investigation of the Thermomechanical Response of a Titanium Aluminide Metal Matrix Composite Using a Viscoplastic Constitutive Theory
210-10MG-098 | Dr. James Sherwood |
| 88 | No Report Submitted (1985)
760-0MG-067 | Dr. Robert Swanson |

89	Optical Profiling of Electric Fields in Layered Structures 210-10MG-071	Dr. Michael Sydor
----	--	-------------------

Volume IV

Armstrong Laboratory

Aerospace Medicine Directorate

90	Confirmation of the Possible Role of Lipopolysaccharide in Expressing an Abelson Murine Leukemia Virus in RAW 264.7 Macrophage Cells 210-10MG-009	Dr. Robert Blystone
91	Effect of Microwave Radiation on Cultured Cells 210-10MG-097	Dr. C. Caudle-Alexander
92	In Vivo Processing of Tetraisopropyl Pyrophosphoramine 210-10MG-083	Dr. James Chambers
93	EMG Analysis of Muscular Fatigue and Recovery Following Alternating Isometric Contractions at Different Levels of Force 210-10MG-014	Dr. Mark Cornwall
94	PCR Analysis of Specific Target Sequence of <u>Mycoplasma hominis</u> and <u>Ureaplasma urealyticum</u> 210-10MG-013	Dr. Vito DelVecchio
95	Studies on Melanocytes and Melanins 210-10MG-133	Dr. Gwendolyn Howze
96	No Report Submitted (1985) 760-0MG-110	Dr. Amir Karimi
97	Robust Filtering of Biological Data 210-10MG-092	Dr. Harold Longbotham
98	No Report Submitted (1985) 760-0MG-101	Dr. James Mrotek

99	Adenosine Modulation of Neurotransmitter Release from Hippocampal Mossy Fiber Synaptosomes 760-7MG-091	Dr. Ralph Peters (1987)
100	Behavioral and Neurochemical Effects of Radiofrequency Electromagnetic Radiation 210-10MG-035	Dr. Raymond Quock
101	An Investigation of Dioxin Half-Life Estimation in Humans Based on Two or More Measurements Per Subject 210-10MG-068	Dr. Ram Tripathi
Crew Systems Directorate		
102	No Report Submitted (1985) 760-0MG-049	Dr. John Flach
103	Degradation of the Renal Peritubular Basement Membrane in Relation to Toxic Nephropathy from Compounds of Military Interest 210-10MG-101	Dr. Thomas Lockwood
104	Parametric Studies of the Breakdown of Total Information Processing Time into During-Display and Post-Display Components 210-10MG-024	Dr. Ethel Marin
105	A Blackboard Architecture for Landmark Identification on 3-Dimensional Surface Images of Human Subjects 210-10MG-077	Dr. Randy Pollack
106	Effect of System Reliability on Probabilistic Inference 760-7MG-094	Dr. Donald Robertson (1987)
107	Stable Grasping with the Utah/MIT Dexterous Robot Hand 210-10MG-034	Dr. Michael Stanisic
108	Articulated Total Body (ATB) "View" Program 210-10MG-053	Dr. Chi-Ming Tang

- | | | |
|-----|---|-----------------|
| 109 | Explorations into the Visual Perceptual Factors
Operating in High-Speed Low-Altitude Turns
210-10MG-105 | Dr. Ebo Tei |
| 110 | No Report Submitted (1985)
760-0MG-071 | Dr. Yin-min Wei |

Human Resources Directorate

- | | | |
|-----|--|------------------------|
| 111 | Computer-Based Training for Complex, Dynamic
Tasks
210-10MG-015 | Dr. Kevin Bennett |
| 112 | Report Not Publishable (1987)
760-7MG-100 | Dr. Ronna Dillon |
| 113 | No Report Submitted (1986)
760-6MG-134 | Dr. Stephen Loy |
| 114 | Advancing User Interface Capabilities in an
Integrated Information Environment: A Fisheye
Browser
210-10MG-110 | Dr. Deborah Mitta |
| 115 | An Intelligent Teacher's Associate for
Network Theory Based on the Heuristic of Polya
760-6MG-032 | Dr. Philip Olivier |
| 116 | An Assessment of the Effects of CONFER:
A Text-Based Intelligent Tutoring System
Designed to Enact Tutorial Conversation and to
Increase a Student's Sense of Intertextuality
210-10MG-003 | Dr. William Smith |
| 117 | The Effect of Student-Instructor Interaction
on Achievement in Computer-Based Training
210-10MG-006 | Dr. Stanley Stephenson |
| 118 | No Report Submitted (1985)
760-0MG-030 | Dr. Christian Wagner |
| 119 | An Evaluation of Stereoscopic and Other Depth
Cues in Computer Display
210-10MG-112 | Mr. John Williamson |
| 120 | New Architectures for WISIWYSWIWSWYS
210-10MG-028 | Dr. Michael Wolfe |

- | | | |
|-----|--|--------------------|
| 121 | Variable Resolution Imagery for Flight Simulators
210-10MG-130 | Dr. Yehoshua Zeevi |
| 122 | Neurocomputing in Intelligent Tutors: Student
Model Diagnosis
210-10MG-063 | Dr. Robert Zerwekh |

Occupational and Environmental Health Directorate

- | | | |
|-----|--|--------------------|
| 123 | Automatic Radiofrequency Radiation Measurement
System
210-10MG-081 | Dr. Stewart Maurer |
|-----|--|--------------------|

Final Report
1990 USAF-UES SUMMER FACULTY RESEARCH PROGRAM/
GRADUATE STUDENT RESEARCH PROGRAM

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the
Universal Energy Systems, Inc.

FINAL REPORT

OPTIMAL DESIGN OF FINITE WORDLENGTH FIR DIGITAL FILTERS
FOR AN ANALOG TRANSVERSAL FILTER WITH TAP WEIGHT
CIRCUITRY DEFECTS USING ADAPTIVE MODELING

Prepared by:	Charles J. Alajajian
Academic Rank:	Assistant Professor
Department and	Electrical and Computer Engineering Dept.
University:	West Virginia University
Research Location:	Rome Labs DCCD Griffis AFB Rome, NY 13441-5700
USAF Researcher:	Richard Hinman
Date:	25 Oct. 91
Contract No:	F49620-88-C-0053 210-10MG-123

Acknowledgements

I would like to thank the Air Force System Command and the Air Force Office of Scientific Research for sponsorship of this research. The assistance of Dr. Rod Darrah, and especially Debbie and Sue of Universal Energy Systems, for their unparalleled courtesy, helpfulness and expert handling of the administrative and directional aspects of the summer program is gratefully acknowledged.

In addition, I would like to personally thank USAF researcher Rick Hinman, of DCCD, for his interest in and support of this research and for presenting me with many challenging theoretical problems. I am grateful to Rick for introducing me to ACT technology and its many applications.

Peter Leong, Director of the Communications Technology Branch, provided a professional and cordial atmosphere in which to work; he is a model branch director and exemplifies what a department head should be.

I would also like to thank Dr. Fred I. Diamond, Chief Scientist at Rome Labs, for meeting with Rick Hinman and me, and for his support of this project.

My wife Hilda, was a constant source of encouragement throughout this work.

Much appreciation to my wonderful parents and the congregation of the South Hills Evangelical Free Church, who have upheld me with their prayers and financial support during this stressful time during which I have been unemployed.

"Ask Yahweh for rain at the time of the spring rains. For it is Yahweh who sends the lightning and gives the showers of rain; he gives bread to man, and grass to cattle."

Zechariah 10: 1 (JB)

Optimal Design Of Finite Wordlength FIR Digital Filters
For an Analog Transversal Filter With Tap Weight
Circuitry Defects Using Adaptive Modeling

by

Charles J. Alajajian

ABSTRACT

An algorithm is presented for designing optimal finite wordlength FIR digital filters for a programmable analog transversal filter with known tap weight circuitry defects. The technique is an application of an adaptive modeling scheme formerly used by Widrow and Stearns for the synthesis of infinite-precision FIR digital filters from specified frequency response characteristics.

Unlike many existing algorithms which utilize successive rounding of the infinite-precision coefficients, the proposed adaptive modeling scheme for the optimal (Chebyshev) design of finite wordlength FIR digital filters incorporates the finite wordlength restriction as part of the filter design procedure. The frequency response characteristics in the fundamental frequency range are specified by taking the DFT of the optimal infinite-precision filter coefficients; these coefficients are obtained via the Parks-McClellan algorithm.

A standard system identification architecture is used in which the optimal infinite-precision filter represents the plant, which is known in this scheme. The plant and the adaptive filter are simultaneously driven by an input signal consisting of a sum of sinusoids. Using the truncated, optimal infinite-precision coefficients as the initial tap weight vector, the LMS algorithm, which is implemented digitally, adjusts the tap weights from among a finite set of fixed-point numbers, in an attempt to minimize the mean square error in the frequency response.

The algorithm is conceptually simple, requires very little computational effort, and is effective even for relatively long filter lengths.

I. INTRODUCTION

The programmable analog transversal filter (PTF) prototype used in this work is based upon an acoustic charge transport (ACT) tapped delay line with active GaAs circuits for coefficient storage and the tap weight circuitry [1]. When a digital filter is implemented on this special-purpose hardware, the infinite-precision filter coefficients are represented internally by the ACT PTF in digital memory by a 6-bit, signed magnitude, tap weight word [1]. This internal quantization results in a departure from the ideal frequency response. Coefficient-quantization errors in FIR digital filters have been thoroughly studied by many authors, among them [5], [15].

In the manufacture of ACT PTF's, the multiple tap weight circuits may exhibit some error in representing the tap weights, primarily due to certain parasitic and nonideal effects unavoidable in a practical weighting circuit. However, by careful modeling, these effects may be compensated for [1].

Of a more severe nature are defects in the tap weight circuitry which preclude the ability to completely address, or alter the tap weight values. That is, one or more bits may be fixed at some prescribed level — fully on, fully off, or at some level in between [18]. These tap weight circuitry defects manifest themselves as errors in the filter's frequency response.

An adaptive modeling scheme formerly used by Widrow and Stearns [12], [13] for the synthesis of FIR digital filters is modified and applied to the design of optimal finite wordlength FIR digital filters for an ACT programmable transversal filter with these tap weight circuitry defects. The modified scheme uses a digital implementation of the LMS algorithm which attempts to find a set of tap weights from among a finite set of fixed-point numbers, which minimizes the mean-squared error in the frequency response.

My research interests have been in the areas of computer-aided design of electronic circuits and in digital signal processing; this provides a wide array of tools with which to address this problem.

II. OBJECTIVES

The primary objective of this research effort is to develop an algorithm to design optimal finite wordlength FIR digital filters for a programmable transversal filter, based upon ACT technology, with tap weight circuitry defects. Computer simulation will be used to assess the effectiveness of the algorithm and to simulate the tap weight circuitry defects of the actual ACT

PTF hardware.

Presently, a prototype analog PTF, built primarily for demonstration purposes [1] consists of an analog tapped-delay line with 64 taps which utilizes digital memory to store the binary representation of each of the desired tap weights; in [1] this is termed a digital/analog (D/A) PTF. The primary components are an analog tapped-delay line and digital reference storage registers and weighting circuits. The analog tapped-delay line processes the input signal; the digital storage registers record the tap weight words; the weighting circuits, consisting of multiplying D/A converters (MDAC) weight the analog signals from the taps according to the digital tap weight words. The prototype D/A PTF referred to in this paper consists of a 64-tap ACT delay line with a SAW frequency of 360 Mhz, 64 MDAC circuits which utilize signed magnitude word format for the tap weights, and a 64X6 static random-access tap weight store, with an address decoder. The tap weight word length is six bits, including the sign bit, so that the smallest number variation that can be represented corresponds to $\frac{1}{32}$, that is, one LSB of the tap weight range [1]. This internal quantization of the filter coefficients, introduces movement in the nominal position of the zeros of the FIR filter, which results in errors in the frequency response.

Nonuniformities in the active devices result in nonuniformities in the tap weights, which are manifest as errors in the frequency response, although ostensibly, the MDAC weighting circuits eliminate these errors to first order. In practice, the MDAC circuits introduce errors in the actual weights due to parasitic capacitances in the C/2C ladder circuits, parasitics in the active switches, and the finite impedances of significant summing buses. Identification and compensation for these errors are feasible if the MDAC performance is carefully modeled [1].

In addition, limitations in tap weight uniformity is due to variations in component parameters among all the MDAC circuits, and to ACT delay line nonuniformities. According to [1], the latter can be compensated for, if a memory-mapping algorithm is used with a small reduction in the overall tap programming range.

III. DESIGN OF FINITE WORDLENGTH FIR DIGITAL FILTERS

a. An adaptive modeling scheme is first utilized to design optimal finite wordlength FIR digital filters assuming no tap weight circuitry defects. The problem is to design an FIR filter, whose coefficients are quantized to a specified finite word length, which best approximates the frequency response of an ideal filter whose coefficients have infinite precision. Other methods have been proposed [2]–[6], [19], but, with the exception of [2], the methods are effective only

for filter lengths of ten or less.

The method proposed in [2] yields the optimal (in the Chebyshev, or min-max sense) finite wordlength FIR filter coefficients using mixed-integer programming methods to solve the discrete optimization problem. Besides its complexity, the method is very costly in terms of the amount of computer time required [2]. However, it is reported to be the only general way of obtaining the optimal finite-wordlength coefficients [2]. The idea behind many of these methods is to obtain the optimal finite-precision coefficients from an iterative process in which the infinite-precision coefficients are successively rounded [2], [19]; the finite-wordlength restriction is not incorporated into the filter design procedure and thus the solution obtained may be suboptimal [2]. However, the proposed adaptive modeling scheme for the optimal design of finite-wordlength FIR digital filters incorporates the finite wordlength restriction as part of the filter design procedure. Although optimality of the coefficients cannot be proved because the finite-precision implementation of the LMS algorithm may deviate from the infinite-precision performance, the method offers noticeable improvement over the truncated infinite-precision coefficients.

This problem of finding a suitable FIR filter with quantized weights may be modeled as an adaptive identification problem and the same basic architecture may be used [12]–[14]. The architecture, depicted in Figure 1, consists of an adaptive filter and a plant which are simultaneously driven by the same input signal $x(k)$. The output of the plant, $d(k)$, supplies the ideal or desired response for the adaptive filter; the subscript k denotes the discrete-time sample index. In the identification problem, the plant is usually unknown and the adaptive filter is used to find a linear model which represents the best fit to the unknown plant; in the case of a dynamic plant the tap weights of the model are time varying [14].

In the present problem, the plant is a known entity; it is the ideal filter whose coefficients have infinite-precision. These coefficients, which constitute the filter's unit-sample response, may be obtained using any FIR filter design technique and are assumed to be exact to within computer accuracy. However, because optimal (Chebyshev) design is the focus of this research, the Parks-McClellan algorithm is exclusively used [9], [16]. The plant is not dynamic since the filter coefficients are not time-varying. The adaptive filter is a finite wordlength FIR filter which attempts to adjust its coefficients to closely approximate the ideal filter response. Because the coefficients are quantized, however, the rule for the updating the adaptive filter weights is implemented digitally; that is, the internal algorithmic calculations are quantized to a finite precision [7], [14]. Thus, the scheme attempts to find the finite-precision model which represents

the best fit to the infinite-precision filter. While in most finite-precision realizations the input samples are also quantized to a finite precision, in the present scheme only the adaptive filter weights are quantized; the samples of the input signal $x(k)$, the ideal, or desired response $d(k)$ and the adaptive filter output $y(k)$ are all assumed to be infinite-precision quantities as implied by Figure 2.

The basic architecture of Figure 1 has been successfully employed by Widrow and Stearns to synthesize infinite-precision FIR filters from a set of design specifications [12], [13]. There are some pronounced differences between the infinite-precision synthesis scheme of Widrow and Stearns and the finite-wordlength design proposed here. In the former scheme, the plant, which is designated as a "pseudo filter," is generally not physically realizable; its unit-sample response is not known; only the output response of the pseudo filter to a sinusoidal input is known. This is to be contrasted with the proposed algorithm for finite-wordlength design in which the plant is a filter whose unit-sample response (which is exact to computer accuracy) is specified; thus, the output of this ideal filter to a given sinusoidal input is also known.

A more salient difference between the two schemes is that the internal calculations are assumed to be of infinite-precision in the former scheme; that is an analog model is assumed; in the latter scheme, the adaptive algorithm is implemented digitally.

The finite-precision form of the LMS algorithm, as it applies to the design of finite wordlength FIR filters is now presented [7], [14]. Referring to Figure 2, the k th output sample of the adaptive filter, $y(k)$, may be computed from

$$y(k) = \mathbf{w}_q^T(k) \mathbf{x}(k) \quad (1)$$

where the character T denotes the transpose operator and $\mathbf{w}_q(k)$ is the tap weight vector, an N -vector of time-varying weights; the subscript q indicates number quantization. That is, $\mathbf{w}_q(k) = [w_{1q}(k), w_{2q}(k), \dots, w_{Nq}(k)]^T$. The tap-input vector $\mathbf{x}(k)$, is an N -vector consisting of the last N samples of the input signal, given by $\mathbf{x}(k) = [x(k), x(k-1), \dots, x(k-N+1)]^T$

The adaptive filter employs the LMS algorithm, which attempts to adjust the weights such that the output of the adaptive filter $y(k)$, and the desired signal $d(k)$ are equal. The error incurred in this process for the k th sample is

$$e(k) = d(k) - y(k) \quad (2)$$

The finite-precision, least-mean square (LMS) algorithm is thus summarized by equation (3) and equation (4) below [7], [14].

$$\epsilon(k) = d(k) - \mathbf{w}_Q^T \mathbf{x}(k) \quad (3)$$

$$\mathbf{w}_Q(k+1) = \mathbf{w}_Q(k) + Q \{ \mu \epsilon(k) \mathbf{x}(k) \} \quad (4)$$

In equation (4), $Q\{ \}$ is an operator denoting an ideal quantizer and μ is a scalar quantity sometimes referred to as the adaptation constant. Only the product representing the gradient vector estimate $\mu \epsilon(k) \mathbf{x}(k)$ is quantized before addition to the tap weight accumulator to form the current tap weight; the starting weight values in equation (4) are the truncated coefficients. This is a reasonable choice, since the optimal finite wordlength coefficients are not expected to deviate very much from the truncated or rounded coefficients [2].

Normally, in a digital implementation, the input signal $\mathbf{x}(k)$, the desired signal $d(k)$, and the internal algorithm are all quantized to a limited precision [14]. However, in this particular application only the internal algorithm for updating the weight vector is quantized, as indicated by equation (4).

The LMS algorithm is known to be an unstable algorithm when implemented digitally [7]. That is, the algorithm will not limit the maximum deviation from infinite-precision performance to within finite bounds, potentially resulting in an overflow. Fortunately, numerical stability is usually not an important property when an adaptive filter is used to determine an unknown setting and then the weights are held fixed at that setting [7]. Since the FIR filter is not a time-varying system, use of the finite-precision LMS algorithm is not precluded.

In the finite-precision implementation, the adaptation constant must be carefully chosen in order to minimize the deviation between the finite-precision and infinite-precision performance. Several authors have studied the choice of the adaptation constant; it is shown in [7]–[8], that increasing the adaptation constant μ minimizes the deviation from infinite-precision performance although decreasing μ below a certain value actually increases the mean-square prediction error. This behavior is different from the infinite-precision case where decreasing μ is known to improve performance when there is no change or a slowly varying relationship between $\mathbf{x}(k)$ and $d(k)$. Compounding the problem of the choice of μ in the finite-precision case is the conflicting fact that increasing μ can also magnify numerical errors. Thus, there is some tradeoff to be made in the choice of μ , which generally involves some measure of “cut and try” [7].

b. The adaptive filter and the ideal filter are simultaneously excited by a sum of M sinusoids, [12]–[13], whose radian frequencies $\omega_i T = 2\pi f_i / f_s$ ($i=1, \dots, M$) consist of uniformly spaced points around the unit circle between $[0, \pi]$ radians. While it is not necessary that the points be uniformly spaced, this is a convenient choice which works well in the present scheme. The sampling period T is normalized to unity for convenience; thus the Nyquist frequency is π r/s, or 0.5 hz. The k th sample of the input signal $x(k)$, is given by

$$x(k) = \sum_{i=1}^M c_i \sin 2\pi f_i k \quad (5)$$

Increasing the amplitude c_i of the i th sinusoid has the effect of more tightly holding the desired response at the i th frequency which might be desirable if unsatisfactory results are obtained [12]. In the present application, setting all of the c_i equal so that each sinusoid has equal amplitude is usually satisfactory.

The signal $x(k)$ must be properly scaled so that the quantizer input, i. e. the quantity in brackets in equation (4), is confined to the interval $[-1, +1]$. Since $x(k)$ is a deterministic (sinusoidal) signal, ensuring that the range of $x(k)$ is confined to this interval is easily accomplished. Initially all the c_i are set according to the weighting desired (in this case unity), and the resulting range of $x(k)$ is observed. If the observed range of $x(k)$ is confined to the interval $[-R, R]$, $x(k)$ is properly scaled by dividing all the c_i by R .

The ideal filter response $d(k)$, at time k , contains the same terms as the input signal, but each individual sinusoid is scaled in amplitude and shifted in phase by amounts a_i and θ_i , respectively, which are the amplitude and phase of the infinite-precision filter measured at the i th frequency. Therefore, $d(k)$ is given by

$$d(k) = \sum_{i=1}^M c_i a_i \sin(2\pi f_i k + \theta_i) \quad (6)$$

If there is to be no error due to overflow, the range of $d(k)$ must also be confined to the interval $[-1, +1]$; this is a common assumption in fixed-point analysis [8]. This is not a problem because the digital filter obtained via the Parks-McClellan algorithm is nonamplifying; that is, the output power is less than or equal to the input power and $|h(n)| \leq 1$ for all n [2].

The triplet (f_l, a_l, θ_l) is seen to characterize the ideal response. This triplet is obtained as follows.

(1) An FIR filter of length- N is designed using the Parks-McClellan algorithm [9], [16], to give the infinite-precision unit-sample response $h(n)$ of length- N . Since optimal Chebyshev design is the focus of this research, the Parks-McClellan algorithm is used. However, any of the many other available design techniques [10], [16], such as frequency-sampling design, least-squared error frequency-domain design, or window-based designs could be used instead.

(2) An L -point FFT of $h(n)$ is taken where $L \geq N$; the relationship between L and N will soon become evident. It is well known that the FFT is simply an efficient method for computing the DFT. The L -point DFT may be interpreted as samples of the frequency response $H(e^{j\omega})$ evaluated at L uniformly spaced frequency points ω on the unit circle, ranging from zero to 2π , or f ranging from zero to unity [16]. That is,

$$H(l) = H(e^{j\omega_l}) \Big|_{\omega_l = \frac{2\pi(l-1)}{L}} \quad 1 \leq l \leq L \quad (7)$$

Clearly, from equations (5) and (6), only M frequencies in the fundamental range $[0, 0.5]$ are of interest. Thus, if L is even, only the first $M = \frac{L}{2} + 1$ complex DFT samples of the L -point DFT are used; the last sample corresponds to the Nyquist frequency. If L is odd, only the first $M = \frac{(L+1)}{2}$ DFT samples are used; note that the last sample does not correspond to the Nyquist frequency. Since the sampling frequency is normalized to unity, the sample (bin) number l and the actual frequency are related by

$$f_l = \frac{l-1}{L} \quad 1 \leq l \leq M \quad (8)$$

where $M = \frac{L}{2} + 1$ (L even) and $M = \frac{(L+1)}{2}$ (L odd) and the bins are numbered starting with unity. The amplitude and phase of $H(l)$ are specified at these M frequencies according to

$$a_l = |H(l)| \quad 1 \leq l \leq M \quad (9)$$

$$\theta_l = \arg H(l) \quad 1 \leq l \leq M \quad (10)$$

If $M = N$ frequency samples are used, which require an $L = (2N-2)$ -point DFT (N even) or an $L = (2N-1)$ -point DFT (N odd), the finite-precision LMS algorithm is observed to render unsatisfactory results. Ostensibly, a larger number of frequency samples is needed to

satisfactorily describe the magnitude and phase characteristics of the ideal filter. By increasing the number of frequency samples to $M=2N$, the algorithm consistently yields very good results. This requires an $L=(4N-2)$ -point DFT (N even) or an $L=(4N-1)$ -point DFT (N odd). Thus, the unit-sample response of the ideal, infinite-precision filter of length N is padded with $3N-2$ zeros (N even) or $3N-1$ zeros (N odd) prior to taking the DFT, yielding $2N$ equally spaced frequency samples between $[0,0.5]$. Because of the particular FFT algorithm used to compute the DFT, the filter length N is not restricted to be a power of 2.

The Parks-McClellan algorithm may force the ideal filter response to zero at either $f=0.0$ or $f=0.5$ [16]. In the latter case, it may be desirable to utilize the odd, $L=(4N-1)$ -point DFT; otherwise, there will automatically be one less sinusoidal term in equation (6).

IV. FUNCTION AND OPERATION OF THE QUANTIZER

a. The ACT PTF cannot internally represent the filter coefficients with infinite-precision or an unlimited number of bits. To model the quantization operation on the computer, each infinite-precision coefficient value (represented by a floating-point number on the computer) is converted into a second, floating-point number equal to the finite-precision value closest to it, that can be represented in a signed magnitude, fixed-point number system having $b-1$ bits (excluding the sign bit). The finite-precision values that can be represented in the number system are called quantization levels; the quantization step Q , is the distance between two adjacent quantization levels; $Q=2^{-(b-1)}$ corresponds to a binary "1" at the least significant register position [11], [15], [17], [21].

The ACT PTF represents the tap weight values using a signed magnitude, fixed-point number representation [1]. The optimum infinite-precision coefficients are assumed to be truncated to a wordlength of 5-bits ($b=6$); this determines the quantizer characteristic used in the computer simulation. Note the magnitude of the error after truncation is generally more severe than after rounding; in the former case it is confined to the interval $[-Q, Q]$ while in the latter case it is confined to the interval $[-\frac{Q}{2}, \frac{Q}{2}]$, [15], [17].

V. MODIFICATION OF THE LMS ALGORITHM TO ENSURE LINEAR PHASE

The LMS algorithm may not always yield filters with linear phase even if the frequency response of the optimal, infinite-precision filter possesses the linear phase property. However, the algorithm may be modified so that the coefficients are adjusted symmetrically, thus ensuring filters with linear phase, as follows [12], [22].

Let $I = \lfloor \frac{N}{2} \rfloor$, denote the number of symmetric weight pairs, where N is the filter length, and $\lfloor x \rfloor$ denotes the largest integer less than or equal to x . For an odd length filter, the middle weight is updated according to

$$w_{(I+1)q}^{k+1} = w_{(I+1)q}^k + \mu e(k)x(k-I) \quad (11)$$

while the lower weights are updated according to

$$w_{(I-l+1)q}^{k+1} = w_{(I-l+1)q}^k + \frac{1}{2}\mu e(k)\{x(k-I+l) + x(k-I-l)\} \quad l=1,2,\dots,I \quad (12)$$

then the upper weights are set equal to $w_{(I+l+1)q}^{k+1} = w_{(I-l+1)q}^{k+1} \quad l=1,2,\dots,I$

Similarly, for an even length filter, the lower weights are updated according to

$$w_{(I-l+1)q}^{k+1} = w_{(I-l+1)q}^k + \frac{1}{2}\mu e(k)\{x(k-I+l) + x(k-I-l+1)\} \quad l=1,2,\dots,I \quad (13)$$

then the upper weights are set equal to $w_{(I+l)q}^{k+1} = w_{(I-l+1)q}^{k+1} \quad l=1,2,\dots,I$

The components of the finite-precision tap weight vector are designated by the q subscript, while the sample index k is indicated by a superscript to simplify the otherwise cumbersome notation.

VI. TESTING OF THE FINITE WORDLENGTH FIR FILTER DESIGN ALGORITHM

a. Six filters of various lengths and types were designed using the Parks—McClellan algorithm [9], [16], to test the effectiveness of finite-wordlength FIR filter design based on adaptive modeling *without* tap weight circuitry defects. Using the truncated coefficients as the starting weights, the corresponding finite-wordlength FIR filters were designed using the proposed adaptive modeling algorithm. The optimal infinite-precision coefficients, the truncated coefficients and the optimal finite-wordlength coefficients obtained using the proposed algorithm are recorded for each test filter, and the magnitude responses are compared.

b. A length-21 Low-pass Filter

FINITE IMPULSE RESPONSE (FIR) LINEAR PHASE DIGITAL FILTER DESIGN REMEX EXCHANGE ALGORITHM

FILTER LENGTH = 21

***** UNIT-SAMPLE RESPONSE *****

h(1) = .18255444E-01 = h(21)
h(2) = .55136724E-01 = h(20)
h(3) = -.40910738E-01 = h(19)
h(4) = .14930869E-01 = h(18)
h(5) = .27568590E-01 = h(17)
h(6) = -.59407810E-01 = h(16)
h(7) = .44841838E-01 = h(15)
h(8) = .31902670E-01 = h(14)
h(9) = -.14972545E+00 = h(13)
h(10) = .25687238E+00 = h(12)
h(11) = .69994063E+00 = h(11)

	Band 1	Band 2
Lower band edge	0.0	0.37
Upper band edge	0.33	0.50
Desired value	1.0	0.0
Weighting	1.0	1.0
Deviation	0.0988697	0.0988697
Deviation in dB	0.8189237	-20.0987384

The truncated, 6-bit coefficients, multiplied by $2^5 = 32$ are

h(1) = 0 = h(21)
h(2) = 1 = h(20)
h(3) = -1 = h(19)
h(4) = 0 = h(18)
h(5) = 0 = h(17)
h(6) = -1 = h(16)
h(7) = 1 = h(15)
h(8) = 1 = h(14)
h(9) = -4 = h(13)
h(10) = 8 = h(12)
h(11) = 22 = h(11)

The optimal, 6-bit coefficients, multiplied by $2^5 = 32$ are

$$\begin{aligned}h(1) &= 0 = h(21) \\h(2) &= 2 = h(20) \\h(3) &= -1 = h(19) \\h(4) &= 0 = h(18) \\h(5) &= 1 = h(17) \\h(6) &= -2 = h(16) \\h(7) &= 1 = h(15) \\h(8) &= 1 = h(14) \\h(9) &= -5 = h(13) \\h(10) &= 8 = h(12) \\h(11) &= 22 = h(11)\end{aligned}$$

It is interesting to note the proximity of the optimal coefficients to the truncated coefficients. A comparison of the magnitude responses is shown in Figure 3. The LMS convergence is noted by plotting the error $e(k)$ versus the iteration number; this is shown in Figure 4.

c. A length-20 Low-pass Filter

FINITE IMPULSE RESPONSE (FIR)
LINEAR PHASE DIGITAL FILTER DESIGN
REMEX EXCHANGE ALGORITHM

FILTER LENGTH = 20

***** UNIT-SAMPLE RESPONSE *****

$$\begin{aligned}h(1) &= .48411223E-01 = h(20) \\h(2) &= .13537383E-01 = h(19) \\h(3) &= -.39344038E-01 = h(18) \\h(4) &= .53151826E-01 = h(17) \\h(5) &= -.31608274E-01 = h(16) \\h(6) &= -.25162720E-01 = h(15) \\h(7) &= .83330645E-01 = h(14) \\h(8) &= -.86372217E-01 = h(13) \\h(9) &= -.34074439E-01 = h(12) \\h(10) &= .56718866E+00 = h(11)\end{aligned}$$

	Band 1	Band 2
Lower band edge	0.0	0.37
Upper band edge	0.33	0.50
Desired value	1.0	0.0
Weighting	1.0	1.0
Deviation	0.0.0981161	0.0981161
Deviation in dB	0.8129651	-20.1651949

Because the frequency response of this filter is forced to be zero at $f=0.5$, an odd $(4N-1)=79$ -point DFT is used for the reason explained in section III.

The truncated, 6-bit coefficients, multiplied by $2^5 = 32$, are

$$\begin{aligned}
 h(1) &= 1 = h(20) \\
 h(2) &= 0 = h(19) \\
 h(3) &= -1 = h(18) \\
 h(4) &= 1 = h(17) \\
 h(5) &= -1 = h(16) \\
 h(6) &= 0 = h(15) \\
 h(7) &= 2 = h(14) \\
 h(8) &= -2 = h(13) \\
 h(9) &= -1 = h(12) \\
 h(10) &= 18 = h(11)
 \end{aligned}$$

The optimal, 6-bit coefficients, multiplied by $2^5 = 32$, are

$$\begin{aligned}
 h(1) &= 1 = h(20) \\
 h(2) &= 0 = h(19) \\
 h(3) &= -1 = h(18) \\
 h(4) &= 2 = h(17) \\
 h(5) &= -1 = h(16) \\
 h(6) &= -1 = h(15) \\
 h(7) &= 2 = h(14) \\
 h(8) &= -3 = h(13) \\
 h(9) &= -1 = h(12) \\
 h(10) &= 18 = h(11)
 \end{aligned}$$

A comparison of the magnitude responses is shown in Figure 5. A plot of the error $e(k)$ versus the iteration number is shown in Figure 6.

d. A length-21 Bandpass Filter

FINITE IMPULSE RESPONSE (FIR)
LINEAR PHASE DIGITAL FILTER DESIGN
REMEX EXCHANGE ALGORITHM

FILTER LENGTH = 21

***** UNIT-SAMPLE RESPONSE *****

h(1) = .46678024E-02 = h(21)
h(2) = .96758919E-02 = h(20)
h(3) = -.90181293E-01 = h(19)
h(4) = -.25750540E-01 = h(18)
h(5) = .45590497E-01 = h(17)
h(6) = -.10308870E-01 = h(16)
h(7) = .11038485E+00 = h(15)
h(8) = .12596292E-01 = h(14)
h(9) = -.28589705E+00 = h(13)
h(10) = -.17343520E-01 = h(12)
h(11) = .38577729E+00 = h(11)

	Band 1	Band 2	Band 3
Lower band edge	0.0	0.18	0.37
Upper band edge	0.14	0.33	0.50
Desired value	0.0	1.0	0.0
Weighting	1.0	1.0	1.0
Deviation	0.1073546	0.1073546	0.1073546
Deviation in dB	-19.3835875	0.8857342	-19.3835875

The truncated, 6-bit coefficients, multiplied by $2^5 = 32$ are

h(1) = 0 = h(21)
h(2) = 0 = h(20)

$$\begin{aligned}
h(3) &= -2 = h(19) \\
h(4) &= 0 = h(18) \\
h(5) &= 1 = h(17) \\
h(6) &= 0 = h(16) \\
h(7) &= 3 = h(15) \\
h(8) &= 0 = h(14) \\
h(9) &= -9 = h(13) \\
h(10) &= 0 = h(12) \\
h(11) &= 12 = h(11)
\end{aligned}$$

The optimal, 6-bit coefficients, multiplied by $2^5 = 32$, are

$$\begin{aligned}
h(1) &= 0 = h(21) \\
h(2) &= 1 = h(20) \\
h(3) &= -3 = h(19) \\
h(4) &= -1 = h(18) \\
h(5) &= 2 = h(17) \\
h(6) &= 0 = h(16) \\
h(7) &= 4 = h(15) \\
h(8) &= 1 = h(14) \\
h(9) &= -9 = h(13) \\
h(10) &= 0 = h(12) \\
h(11) &= 13 = h(11)
\end{aligned}$$

A comparison of the magnitude responses is shown in Figure 7. A plot of the error $e(k)$ versus the iteration number is shown in Figure 8.

e. A length-32 (Wide-band) Bandpass Filter

FINITE IMPULSE RESPONSE (FIR)
 LINEAR PHASE DIGITAL FILTER DESIGN
 REMES EXCHANGE ALGORITHM

FILTER LENGTH = 32
 ***** UNIT-SAMPLE RESPONSE *****

$h(1) = -.48581465E-02 = h(32)$
 $h(2) = -.84469004E-02 = h(31)$
 $h(3) = .72525650E-02 = h(30)$
 $h(4) = .13274785E-02 = h(29)$
 $h(5) = .16371012E-01 = h(28)$
 $h(6) = .10639032E-01 = h(27)$
 $h(7) = -.29763351E-01 = h(26)$
 $h(8) = -.31177921E-02 = h(25)$
 $h(9) = -.36395655E-01 = h(24)$
 $h(10) = .74851950E-02 = h(23)$
 $h(11) = .82077622E-01 = h(22)$
 $h(12) = -.81893674E-02 = h(21)$
 $h(13) = .73856838E-01 = h(20)$
 $h(14) = -.11670402E+00 = h(19)$
 $h(15) = -.30804001E+00 = h(18)$
 $h(16) = .31369272E+00 = h(17)$

	Band 1	Band 2	Band 3
Lower band edge	0.0	0.20	0.4250
Upper band edge	0.1	0.35	0.50
Desired value	0.0	1.0	0.0
Weighting	1.0	1.0	1.0
Deviation	0.0056256	0.0056256	0.0056256
Deviation in dB	-44.9966866	0.0487261	-44.9966866

The truncated, 6-bit coefficients, multiplied by $2^5 = 32$ are

$h(1) = 0 = h(32)$
 $h(2) = 0 = h(31)$
 $h(3) = 0 = h(30)$
 $h(4) = 0 = h(29)$
 $h(5) = 0 = h(28)$
 $h(6) = 0 = h(27)$
 $h(7) = 0 = h(26)$
 $h(8) = 0 = h(25)$
 $h(9) = -1 = h(24)$

$$\begin{aligned}
h(10) &= 0 = h(23) \\
h(11) &= 2 = h(22) \\
h(12) &= 0 = h(21) \\
h(13) &= 2 = h(20) \\
h(14) &= -3 = h(19) \\
h(15) &= -9 = h(18) \\
h(16) &= 10 = h(17)
\end{aligned}$$

The optimal, 6-bit coefficients, multiplied by $2^5 = 32$, are

$$\begin{aligned}
h(1) &= 0 = h(32) \\
h(2) &= 0 = h(31) \\
h(3) &= 0 = h(30) \\
h(4) &= 0 = h(29) \\
h(5) &= 0 = h(28) \\
h(6) &= 0 = h(27) \\
h(7) &= 0 = h(26) \\
h(8) &= 0 = h(25) \\
h(9) &= -1 = h(24) \\
h(10) &= 0 = h(23) \\
h(11) &= 2 = h(22) \\
h(12) &= 0 = h(21) \\
h(13) &= 3 = h(20) \\
h(14) &= -4 = h(19) \\
h(15) &= -11 = h(18) \\
h(16) &= 11 = h(17)
\end{aligned}$$

While the algorithm yields some improvement as seen from Figure 9, better results might be anticipated if the 10 truncated coefficients with value zero were adjusted to a non-zero value by the adaptive algorithm. A comparison of the magnitude responses is shown in Figure 9. A plot of the error $e(k)$ versus the iteration number is shown in Figure 10.

f. A length-64 (Narrow-band) Bandpass Filter

FINITE IMPULSE RESPONSE (FIR) LINEAR PHASE DIGITAL FILTER DESIGN

REMEX EXCHANGE ALGORITHM

FILTER LENGTH = 64

***** UNIT-SAMPLE RESPONSE *****

h(1) = -.63719952E-02 = h(64)
h(2) = .12366605E-01 = h(63)
h(3) = .90891666E-02 = h(62)
h(4) = -.52737236E-02 = h(61)
h(5) = -.65397498E-02 = h(60)
h(6) = .73926853E-02 = h(59)
h(7) = .72776156E-02 = h(58)
h(8) = -.66523440E-02 = h(57)
h(9) = -.60031978E-02 = h(56)
h(10) = .48160733E-02 = h(55)
h(11) = .32913051E-02 = h(54)
h(12) = -.12530101E-02 = h(53)
h(13) = .11994049E-02 = h(52)
h(14) = -.40770610E-02 = h(51)
h(15) = -.74582934E-02 = h(50)
h(16) = .11142890E-01 = h(49)
h(17) = .15321170E-01 = h(48)
h(18) = -.19652596E-01 = h(47)
h(19) = -.24457549E-01 = h(46)
h(20) = .29206388E-01 = h(45)
h(21) = .34321159E-01 = h(44)
h(22) = -.39200157E-01 = h(43)
h(23) = -.44286929E-01 = h(42)
h(24) = .48936060E-01 = h(41)
h(25) = .53609453E-01 = h(40)
h(26) = -.57691106E-01 = h(39)
h(27) = -.61579415E-01 = h(38)
h(28) = .64760943E-01 = h(37)
h(29) = .67547088E-01 = h(36)
h(30) = -.69580477E-01 = h(35)
h(31) = -.71032365E-01 = h(34)
h(32) = .71727105E-01 = h(33)

	Band 1	Band 2	Band 3
Lower band edge	0.0	0.2375	0.2875
Upper band edge	0.2125	0.2625	0.50
Desired value	0.0	1.0	0.0
Weighting	1.0	1.0	1.0
Deviation	0.0217903	0.0217903	0.0217903
Deviation in dB	-33.2347400	0.1872354	-33.2347400

The truncated, 6-bit coefficients, multiplied by $2^5 = 32$ are

$h(1) = 0 = h(64)$
 $h(2) = 0 = h(63)$
 $h(3) = 0 = h(62)$
 $h(4) = 0 = h(61)$
 $h(5) = 0 = h(60)$
 $h(6) = 0 = h(59)$
 $h(7) = 0 = h(58)$
 $h(8) = 0 = h(57)$
 $h(9) = 0 = h(56)$
 $h(10) = 0 = h(55)$
 $h(11) = 0 = h(54)$
 $h(12) = 0 = h(53)$
 $h(13) = 0 = h(52)$
 $h(14) = 0 = h(51)$
 $h(15) = 0 = h(50)$
 $h(16) = 0 = h(49)$
 $h(17) = 0 = h(48)$
 $h(18) = 0 = h(47)$
 $h(19) = 0 = h(46)$
 $h(20) = 0 = h(45)$
 $h(21) = 1 = h(44)$
 $h(22) = -1 = h(43)$
 $h(23) = -1 = h(42)$
 $h(24) = 1 = h(41)$
 $h(25) = 1 = h(40)$
 $h(26) = -1 = h(39)$

$$\begin{aligned}
h(27) &= -1 = h(38) \\
h(28) &= 2 = h(37) \\
h(29) &= 2 = h(36) \\
h(30) &= -2 = h(35) \\
h(31) &= -2 = h(34) \\
h(32) &= 2 = h(33)
\end{aligned}$$

The optimal, 6-bit coefficients, multiplied by $2^5 = 32$, are

$$\begin{aligned}
h(1) &= 0 = h(64) \\
h(2) &= 0 = h(63) \\
h(3) &= 0 = h(62) \\
h(4) &= 0 = h(61) \\
h(5) &= 0 = h(60) \\
h(6) &= 0 = h(59) \\
h(7) &= 0 = h(58) \\
h(8) &= 0 = h(57) \\
h(9) &= 0 = h(56) \\
h(10) &= 0 = h(55) \\
h(11) &= 0 = h(54) \\
h(12) &= 0 = h(53) \\
h(13) &= 0 = h(52) \\
h(14) &= 0 = h(51) \\
h(15) &= 0 = h(50) \\
h(16) &= 1 = h(49) \\
h(17) &= 1 = h(48) \\
h(18) &= -1 = h(47) \\
h(19) &= -1 = h(46) \\
h(20) &= 1 = h(45) \\
h(21) &= 1 = h(44) \\
h(22) &= -1 = h(43) \\
h(23) &= -1 = h(42) \\
h(24) &= 2 = h(41) \\
h(25) &= 2 = h(40) \\
h(26) &= -2 = h(39) \\
h(27) &= -2 = h(38)
\end{aligned}$$

$$\begin{aligned}
h(28) &= 2 = h(37) \\
h(29) &= 2 = h(36) \\
h(30) &= -2 = h(35) \\
h(31) &= -2 = h(34) \\
h(32) &= 2 = h(33)
\end{aligned}$$

In this example, the adaptive algorithm does a much better job of dealing with the zero coefficients than in the wideband length-32 bandpass filter case. Indeed, in the present case, 10 truncated coefficients which are zero in value, are adjusted by the algorithm to take on a non-zero value. A comparison of the magnitude responses is shown in Figure 11. A plot of the error $e(k)$ versus the iteration number is shown in Figure 12.

g. A length-31 Bandreject Filter

FINITE IMPULSE RESPONSE (FIR)
 LINEAR PHASE DIGITAL FILTER DESIGN
 REMES EXCHANGE ALGORITHM

FILTER LENGTH = 31

***** UNIT-SAMPLE RESPONSE *****

$$\begin{aligned}
h(1) &= .55182726E-05 = h(31) \\
h(2) &= -.26521932E-01 = h(30) \\
h(3) &= .35114622E-04 = h(29) \\
h(4) &= -.10156000E-04 = h(28) \\
h(5) &= -.22336191E-04 = h(27) \\
h(6) &= .44154967E-01 = h(26) \\
h(7) &= -.66720404E-04 = h(25) \\
h(8) &= .45685611E-04 = h(24) \\
h(9) &= -.68394425E-05 = h(23) \\
h(10) &= -.93469017E-01 = h(22) \\
h(11) &= .97577121E-04 = h(21) \\
h(12) &= -.82618001E-04 = h(20) \\
h(13) &= .35329113E-04 = h(19) \\
h(14) &= .31394833E+00 = h(18) \\
h(15) &= -.77643091E-04 = h(17) \\
h(16) &= .50010198E+00 = h(16)
\end{aligned}$$

	Band 1	Band 2	Band 3
Lower band edge	0.0	0.15	0.42
Upper band edge	0.1	0.35	0.50
Desired value	1.0	0.0	1.0
Weighting	1.0	1.0	1.0
Deviation	0.0237675	0.0237675	0.0237675
Deviation in dB	0.2040268	-32.4803274	0.2040268

The truncated, 6-bit coefficients, multiplied by $2^5 = 32$ are

$h(1) = 0 = h(31)$
 $h(2) = 0 = h(30)$
 $h(3) = 0 = h(29)$
 $h(4) = 0 = h(28)$
 $h(5) = 0 = h(27)$
 $h(6) = 1 = h(26)$
 $h(7) = 0 = h(25)$
 $h(8) = 0 = h(24)$
 $h(9) = 0 = h(23)$
 $h(10) = -2 = h(22)$
 $h(11) = 0 = h(21)$
 $h(12) = 0 = h(20)$
 $h(13) = 0 = h(19)$
 $h(14) = 10 = h(18)$
 $h(15) = 0 = h(17)$
 $h(16) = 16 = h(16)$

The optimal, 6-bit coefficients, multiplied by $2^5 = 32$, are

$h(1) = 0 = h(31)$
 $h(2) = -1 = h(30)$
 $h(3) = 0 = h(29)$
 $h(4) = 0 = h(28)$
 $h(5) = 0 = h(27)$
 $h(6) = 1 = h(26)$
 $h(7) = 0 = h(25)$

$$\begin{aligned}
h(8) &= -1 = h(24) \\
h(9) &= 0 = h(23) \\
h(10) &= -4 = h(22) \\
h(11) &= 0 = h(21) \\
h(12) &= 0 = h(20) \\
h(13) &= 0 = h(19) \\
h(14) &= 10 = h(18) \\
h(15) &= 0 = h(17) \\
h(16) &= 16 = h(16)
\end{aligned}$$

A comparison of the magnitude responses is shown in Figure 13. A plot of the error $e(k)$ versus the iteration number is shown in Figure 14.

VII. DESIGN OF FINITE WORDLENGTH FIR DIGITAL FILTERS WITH TAP WEIGHT CIRCUITRY DEFECTS

a. The adaptive modeling scheme may also be utilized to design optimal finite- wordlength FIR filters for the ACT analog transversal filter with tap weight circuitry defects. Because of these defects, the ACT hardware may inaccurately represent the optimal finite-wordlength coefficient values. The same algorithm used to design optimal finite-precision FIR filters (assuming no tap-weight circuitry defects) may be used, with the following modifications. Equations (3) and (4) of the finite-precision least-mean square algorithm must be modified according to

$$\epsilon(k) = d(k) - \mathbf{w}_d^T \mathbf{x}(k) \quad (14)$$

$$\mathbf{w}_d(k+1) = \mathbf{w}_d(k) + D \left\{ Q \left\{ \mu \epsilon(k) \mathbf{x}(k) \right\} \right\} \quad (15)$$

For computer simulation, equation (15) is interpreted as follows. The operator $Q\{ \}$ transforms the infinite-precision filter coefficients (represented by floating point numbers on the computer) into a second set of floating point numbers (the quantized coefficients), equal in value to the quantization level closest to the infinite-precision coefficient value. The defect operator $D\{ \}$ transforms these quantized coefficients into a third set of floating point numbers which are the quantized coefficient values actually seen by the hardware. These are dubbed the defective coefficients, and are denoted by the subscript d . The defective ACT hardware has the effect of merely shifting some of the quantized coefficient values from one quantization level to an entirely different level; the quantization step does not change.

b. If the tap weight circuitry defects in the ACT programmable transversal filter occur symmetrically, so as not to negate the symmetry of the filter coefficients as represented internally by the ACT hardware, then equations (11), (12) or (13) can be employed in lieu of equation (15) to ensure linear phase. The Q and D operators should be introduced in these equations in the same manner as in equation (15) with the starting weights set equal to the defective coefficients.

c: Table 1 below lists the known defects in the ACT hardware [18]. Tap weights are numbered starting with unity; the first bit denotes the most significant bit (MSB), the fifth bit denotes the least significant bit; the sixth bit is reserved for the sign bit.

TABLE 1: Defective Tap Weights in the ACT [18]

<u>Tap #</u>	<u>Bit #</u>	<u>Fault</u>
35	5 (LSB)	unknown
36	2 (MSB-1)	always off
38	1 (MSB)	always on
57	4 (MSB-3)	doesn't turn completely on
64	1 (MSB)	always off
2n for n=1,...,32	5 (LSB)	always off

Note that for filters with odd length (less than length 35), symmetry is always maintained; however, because of the last defect in Table 1, symmetry is not maintained for any even-length filter. Since all that is known about tap weight 57 is that it fails to turn on completely, it is assumed to turn on halfway in the computer model.

VIII. TESTING OF THE FINITE WORDLENGTH FIR FILTER DESIGN ALGORITHM WITH TAP WEIGHT CIRCUITRY DEFECTS

a. The test examples are now used to assess the effectiveness of the proposed adaptive modeling technique for the cancellation of errors due to tap weight circuitry defects in the ACT programmable transversal filter.

b. A length-21 Low-pass Filter

The truncated, 6-bit coefficients, as represented internally by the ACT transversal filter with its tap weight circuitry defects, multiplied by $2^5 = 32$ are

$$\begin{aligned}
h(1) &= 0 = h(21) \\
h(2) &= 0 = h(20) \\
h(3) &= -1 = h(19) \\
h(4) &= 0 = h(18) \\
h(5) &= 0 = h(17) \\
h(6) &= 0 = h(16) \\
h(7) &= 1 = h(15) \\
h(8) &= 0 = h(14) \\
h(9) &= -4 = h(13) \\
h(10) &= 8 = h(12) \\
h(11) &= 22 = h(11)
\end{aligned}$$

The optimal, 6-bit coefficients, obtained using the adaptive algorithm in the presence of tap weight circuitry defects, multiplied by $2^5 = 32$ are

$$\begin{aligned}
h(1) &= 1 = h(21) \\
h(2) &= 2 = h(20) \\
h(3) &= -1 = h(19) \\
h(4) &= 0 = h(18) \\
h(5) &= 1 = h(17) \\
h(6) &= -2 = h(16) \\
h(7) &= 1 = h(15) \\
h(8) &= 0 = h(14) \\
h(9) &= -5 = h(13) \\
h(10) &= 8 = h(12) \\
h(11) &= 22 = h(11)
\end{aligned}$$

A comparison of the magnitude responses is shown in Figure 15. A plot of the error $e(k)$ versus the iteration number is shown in Figure 16.

c. A length-20 Low-pass Filter

The truncated, 6-bit coefficients, as represented internally by the ACT transversal filter with its tap weight circuitry defects, multiplied by $2^5 = 32$ are

$$\begin{aligned}
h(1) &= 1 \\
h(2) &= 0
\end{aligned}$$

$$\begin{aligned}
h(3) &= -1 \\
h(4) &= 0 \\
h(5) &= -1 \\
h(6) &= 0 \\
h(7) &= 2 \\
h(8) &= -2 \\
h(9) &= -1 \\
h(10) &= 18 \\
h(11) &= 18 \\
h(12) &= 0 \\
h(13) &= -2 \\
h(14) &= 2 \\
h(15) &= 0 \\
h(16) &= 0 \\
h(17) &= 1 \\
h(18) &= 0 \\
h(19) &= 0 \\
h(20) &= 0
\end{aligned}$$

Since the defects in the tap weight circuitry are such that the symmetry of the filter coefficients is negated, strictly speaking equations (11), (12) or (13) no longer apply. Yet because these equations are necessary to ensure linear phase filters they are used just the same, although the finite wordlength coefficients obtained using them will not be optimal. These 6-bit coefficients, obtained using the adaptive algorithm in the presence of tap weight circuitry defects, multiplied by $2^5 = 32$ are

$$\begin{aligned}
h(1) &= 2 = h(20) \\
h(2) &= 1 = h(19) \\
h(3) &= -2 = h(18) \\
h(4) &= 2 = h(17) \\
h(5) &= 0 = h(16) \\
h(6) &= -1 = h(15) \\
h(7) &= 2 = h(14) \\
h(8) &= -3 = h(13) \\
h(9) &= -2 = h(12) \\
h(10) &= 18 = h(11)
\end{aligned}$$

A comparison of the magnitude responses is shown in Figure 17. A plot of the error $e(k)$ versus the iteration number is shown in Figure 18.

d. A length-21 Bandpass Filter

The optimal coefficient values of this particular filter are unaffected by the tap weight circuitry defects; therefore, it is not necessary to redesign the filter taking into account tap weight circuitry defects.

e. A length-32 (Wide-band) Bandpass Filter

The truncated, 6-bit coefficients, as represented internally by the ACT transversal filter with its tap weight circuitry defects, multiplied by $2^5 = 32$ are

$h(1) = 0$
 $h(2) = 0$
 $h(3) = 0$
 $h(4) = 0$
 $h(5) = 0$
 $h(6) = 0$
 $h(7) = 0$
 $h(8) = 0$
 $h(9) = -1$
 $h(10) = 0$
 $h(11) = 2$
 $h(12) = 0$
 $h(13) = 2$
 $h(14) = -2$
 $h(15) = -9$
 $h(16) = 10$
 $h(17) = 10$
 $h(18) = -8$
 $h(19) = -3$
 $h(20) = 2$
 $h(21) = 0$
 $h(22) = 2$
 $h(23) = 0$
 $h(24) = 0$

$$\begin{aligned}
h(25) &= 0 \\
h(26) &= 0 \\
h(27) &= 0 \\
h(28) &= 0 \\
h(29) &= 0 \\
h(30) &= 0 \\
h(31) &= 0 \\
h(32) &= 0
\end{aligned}$$

Because the filter length is even, defects in the tap weight circuitry are such that the symmetry of the filter coefficients is negated, strictly speaking equations (11), (12) or (13) no longer apply. Yet because these equations are necessary to ensure linear phase filters they are used just the same, although the finite wordlength coefficients obtained using them will not be optimal. These 6-bit coefficients, obtained using the adaptive algorithm in the presence of tap weight circuitry defects, multiplied by $2^5 = 32$ are

$$\begin{aligned}
h(1) &= 0 &= h(32) \\
h(2) &= 0 &= h(31) \\
h(3) &= 0 &= h(30) \\
h(4) &= 0 &= h(29) \\
h(5) &= 0 &= h(28) \\
h(6) &= 0 &= h(27) \\
h(7) &= 0 &= h(26) \\
h(8) &= 0 &= h(25) \\
h(9) &= 0 &= h(24) \\
h(10) &= 0 &= h(23) \\
h(11) &= 2 &= h(22) \\
h(12) &= 0 &= h(21) \\
h(13) &= 4 &= h(20) \\
h(14) &= -3 &= h(19) \\
h(15) &= -10 &= h(18) \\
h(16) &= 12 &= h(17)
\end{aligned}$$

A comparison of the magnitude responses is shown in Figure 19. A plot of the error $e(k)$ versus the iteration number is shown in Figure 20.

f. A length-64 (Narrow-band) Bandpass Filter

The truncated, 6-bit coefficients, as represented internally by the ACT transversal filter with its tap weight circuitry defects, multiplied by $2^5 = 32$

are

$h(1) = 0$
 $h(2) = 0$
 $h(3) = 0$
 $h(4) = 0$
 $h(5) = 0$
 $h(6) = 0$
 $h(7) = 0$
 $h(8) = 0$
 $h(9) = 0$
 $h(10) = 0$
 $h(11) = 0$
 $h(12) = 0$
 $h(13) = 0$
 $h(14) = 0$
 $h(15) = 0$
 $h(16) = 0$
 $h(17) = 0$
 $h(18) = 0$
 $h(19) = 0$
 $h(20) = 0$
 $h(21) = 1$
 $h(22) = 0$
 $h(23) = -1$
 $h(24) = 0$
 $h(25) = 1$
 $h(26) = 0$
 $h(27) = -1$
 $h(28) = 2$
 $h(29) = 2$
 $h(30) = -2$
 $h(31) = -2$
 $h(32) = 2$

$h(33) = 2$
 $h(34) = -2$
 $h(35) = -2$
 $h(36) = 2$
 $h(37) = 2$
 $h(38) = -17$
 $h(39) = -1$
 $h(40) = 1$
 $h(41) = 1$
 $h(42) = -1$
 $h(43) = -1$
 $h(44) = 1$
 $h(45) = 0$
 $h(46) = 0$
 $h(47) = 0$
 $h(48) = 0$
 $h(49) = 0$
 $h(50) = 0$
 $h(51) = 0$
 $h(52) = 0$
 $h(53) = 0$
 $h(54) = 0$
 $h(55) = 0$
 $h(56) = 0$
 $h(57) = 0$
 $h(58) = 0$
 $h(59) = 0$
 $h(60) = 0$
 $h(61) = 0$
 $h(62) = 0$
 $h(63) = 0$
 $h(64) = 0$

Since the defects in the tap weight circuitry are such that the symmetry of the filter coefficients is negated, strictly speaking, equations (11), (12) or (13) no longer apply. Yet because these equations are necessary to ensure linear phase filters they are used just the same, although the

finite wordlength coefficients obtained using them will not be optimal. These 6-bit coefficients, obtained using the adaptive algorithm in the presence of tap weight circuitry defects, multiplied by $2^5 = 32$ are

$h(1)$	$= -7$	$= h(64)$
$h(2)$	$= 2$	$= h(63)$
$h(3)$	$= 4$	$= h(62)$
$h(4)$	$= 4$	$= h(61)$
$h(5)$	$= -13$	$= h(60)$
$h(6)$	$= 8$	$= h(59)$
$h(7)$	$= 3$	$= h(58)$
$h(8)$	$= -18$	$= h(57)$
$h(9)$	$= 12$	$= h(56)$
$h(10)$	$= -1$	$= h(55)$
$h(11)$	$= -5$	$= h(54)$
$h(12)$	$= -17$	$= h(53)$
$h(13)$	$= 10$	$= h(52)$
$h(14)$	$= -12$	$= h(51)$
$h(15)$	$= 10$	$= h(50)$
$h(16)$	$= 12$	$= h(49)$
$h(17)$	$= 20$	$= h(48)$
$h(18)$	$= -22$	$= h(47)$
$h(19)$	$= 31$	$= h(46)$
$h(20)$	$= -24$	$= h(45)$
$h(21)$	$= 31$	$= h(44)$
$h(22)$	$= 12$	$= h(43)$
$h(23)$	$= -2$	$= h(42)$
$h(24)$	$= -27$	$= h(41)$
$h(25)$	$= 30$	$= h(40)$
$h(26)$	$= 2$	$= h(39)$
$h(27)$	$= 31$	$= h(38)$
$h(28)$	$= 9$	$= h(37)$
$h(29)$	$= 7$	$= h(36)$
$h(30)$	$= -24$	$= h(35)$
$h(31)$	$= -8$	$= h(34)$
$h(32)$	$= -24$	$= h(33)$

A comparison of the magnitude responses is shown in Figure 21. A plot of the error $e(k)$ versus the iteration number is shown in Figure 22.

g. A length-31 Bandreject Filter

The truncated, 6-bit coefficients, as represented internally by the ACT transversal filter with its tap weight circuitry defects, multiplied by $2^5 = 32$ are

$$\begin{aligned}
 h(1) &= 0 = h(31) \\
 h(2) &= 0 = h(30) \\
 h(3) &= 0 = h(29) \\
 h(4) &= 0 = h(28) \\
 h(5) &= 0 = h(27) \\
 h(6) &= 0 = h(26) \\
 h(7) &= 0 = h(25) \\
 h(8) &= 0 = h(24) \\
 h(9) &= 0 = h(23) \\
 h(10) &= -2 = h(22) \\
 h(11) &= 0 = h(21) \\
 h(12) &= 0 = h(20) \\
 h(13) &= 0 = h(19) \\
 h(14) &= 10 = h(18) \\
 h(15) &= 0 = h(17) \\
 h(16) &= 16 = h(16)
 \end{aligned}$$

The optimal, 6-bit coefficients, obtained using the adaptive algorithm in the presence of tap weight circuitry defects, multiplied by $2^5 = 32$ are

$$\begin{aligned}
 h(1) &= 0 = h(31) \\
 h(2) &= 0 = h(30) \\
 h(3) &= 0 = h(29) \\
 h(4) &= 0 = h(28) \\
 h(5) &= 0 = h(27) \\
 h(6) &= 0 = h(26) \\
 h(7) &= 0 = h(25) \\
 h(8) &= 0 = h(24) \\
 h(9) &= 0 = h(23)
 \end{aligned}$$

$$\begin{aligned}
h(10) &= -6 = h(22) \\
h(11) &= 0 = h(21) \\
h(12) &= 0 = h(20) \\
h(13) &= 0 = h(19) \\
h(14) &= 10 = h(18) \\
h(15) &= 0 = h(17) \\
h(16) &= 16 = h(16)
\end{aligned}$$

A comparison of the magnitude responses is shown in Figure 23. A plot of the error $e(k)$ versus the iteration number is shown in Figure 24.

VIII. RECOMMENDATIONS:

a. In this work, an application of an adaptive modeling scheme (originally proposed by Stearns and Widrow [12]–[13] for the synthesis of FIR filters), is implemented digitally to design optimal, finite wordlength FIR filters. The algorithm is further modified to design optimal, finite wordlength FIR filters for an ACT, programmable analog transversal filter with known tap weight circuitry defects. However, in this case, the method is most effective only if the tap weight defects occur symmetrically so as not to negate the symmetry of the filter coefficients as represented internally by the ACT hardware. Furthermore, the digital algorithm may be used to design optimal finite wordlength FIR digital filters directly from a specified set of frequency response characteristics.

Although the method of mixed-integer programming is the only general way reported in the literature for optimal finite-wordlength coefficients [2], the computational efficiency of the proposed algorithm renders it useful for filter lengths that would ordinarily be precluded with the former method. Indeed, the bulk of this work was done using only a desktop PC–XT clone, equipped with a math coprocessor chip.

b. The present work has been restricted to the design of finite wordlength FIR digital filters. It is recommended that the work be extended to include the design of stable IIR filters with finite wordlength. Coefficient quantization is a far more serious problem in IIR filters, since an IIR filter with infinite-precision coefficients, which is stable by design, may actually manifest itself as an unstable filter when implemented in special purpose hardware. Of course, in both FIR and IIR filters, coefficient quantization may result in appreciable deviation in the frequency response from that obtained with infinite-precision coefficients [17].

c. It is anticipated that the performance of the proposed algorithm would be much improved if the number of bits used to represent the infinite-precision coefficients were increased. This would probably reduce the number of "zero tap weights" that are sometimes encountered when using the algorithm.

Further improvement in performance might be realized if the single adaptive linear combiner employed in the present work is replaced by a neural network.

REFERENCES

Conference and journal publications:

1. R. W. Miller, C. A. Ricci and R. J. Kansy, "An Acoustic Charge Transport Digitally Programmable Transversal Filter," IEEE Journal Solid-State Circuits, Vol. 24 Dec. 1989, pp. 1675 - 1682.
2. D. M. Kodek, "Design of Optimal Finite Wordlength FIR Digital Filters Using Integer Programming Techniques," IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-28, June 1980, pp. 304 - 308.
3. M. Suk and S. K. Mitra, "Computer-Aided Design of Digital Filters with Finite Word Lengths," IEEE Trans. Audio and Electroacoustics, Vol AU-20, Dec. 1972, pp. 356 - 363.
4. E. Avenhaus, "On the Design of Digital Filters with Coefficients of Limited Word Length," IEEE Trans. on Audio and Electroacoustics, Vol. AU-20, Aug. 1972, pp. 206 - 212.
5. D. S. K. Chan and L. R. Rabiner, "Analysis of Quantization Errors in the Direct Form for Finite Impulse Response Digital Filters," IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, Aug. 1973, pp. 354 - 366.
6. C. Charalambos and M. J. Best, "Optimization of Recursive Digital Filter with Finite Wordlength," IEEE Trans. Acoust. Speech Signal Process., vol. ASSP-22, , Dec. 1974, pp.424-431.
7. J. M. Cioffi, "Limited-Precision Effects in Adaptive Filtering," IEEE Trans. Circuits and Syst., Vol. CAS-34, July 1987, pp. 821 - 833.
8. S. T. Alexander, "Transient Weight Misadjustment Properties for the Finite Precision LMS Algorithm," IEEE Trans. Acoust. Speech Signal Process., vol. ASSP-35, pp. 1250 - 1258.
9. L. R. Rabiner, J. H. McClellan and T. W. Parks, "FIR Digital Filter Design Techniques Using Weighted Chebyshev Approximation," Proc. IEEE, Vol. 63, April 1975, pp. 595 - 610.
10. O. Herrmann, "On the Approximation Problem in Nonrecursive Digital Filter Design," IEEE Trans. Circuit Theory, Vol. 18, May 1971, pp. 411 - 413.

11. S. K. Tewksbury *et al.*, "Terminology related to the performance of S/H, A/D, and D/A circuits," IEEE Trans. Circuits Syst., Vol. CAS-25, July 1978, pp. 419-426.

Textbooks:

12. B. Widrow and S. D. Stearns, Adaptive Signal Processing, Englewood Cliffs, NJ, Prentice-Hall, 1985.

13. S. D. Stearns and R. A. David, Signal Processing Algorithms, Englewood Cliffs, NJ, Prentice-Hall, 1988.

14. S. Haykin, Adaptive Filter Theory 2nd. Ed. Englewood Cliffs, NJ, Prentice-Hall, 1991.

15. A. Antoniou, Digital Filters: Analysis and Design, New York, NY, Mc-Graw-Hill, 1979.

16. T. W. Parks and C. S. Burrus, Digital Filter Design, New York, NY, Wiley Interscience, 1987.

17. N. K. Bose, Digital Filters, , New York, NY, Elsevier Science Publishing, 1985.

18. Richard Hinman, Personal communication, RADC/DCCD, Rome, NY.

19. U. Heute, "A Subroutine for Finite Wordlength FIR Filter Design," in Programs for Digital Signal Processing, New York, NY, IEEE Press, 1979.

21. R. Kuc, Introduction to Digital Signal Processing, New York, NY, McGraw-Hill, 1988.

22. Samuel D. Stearns, Personal communication, Sandia Labs, Albuquerque, NM.

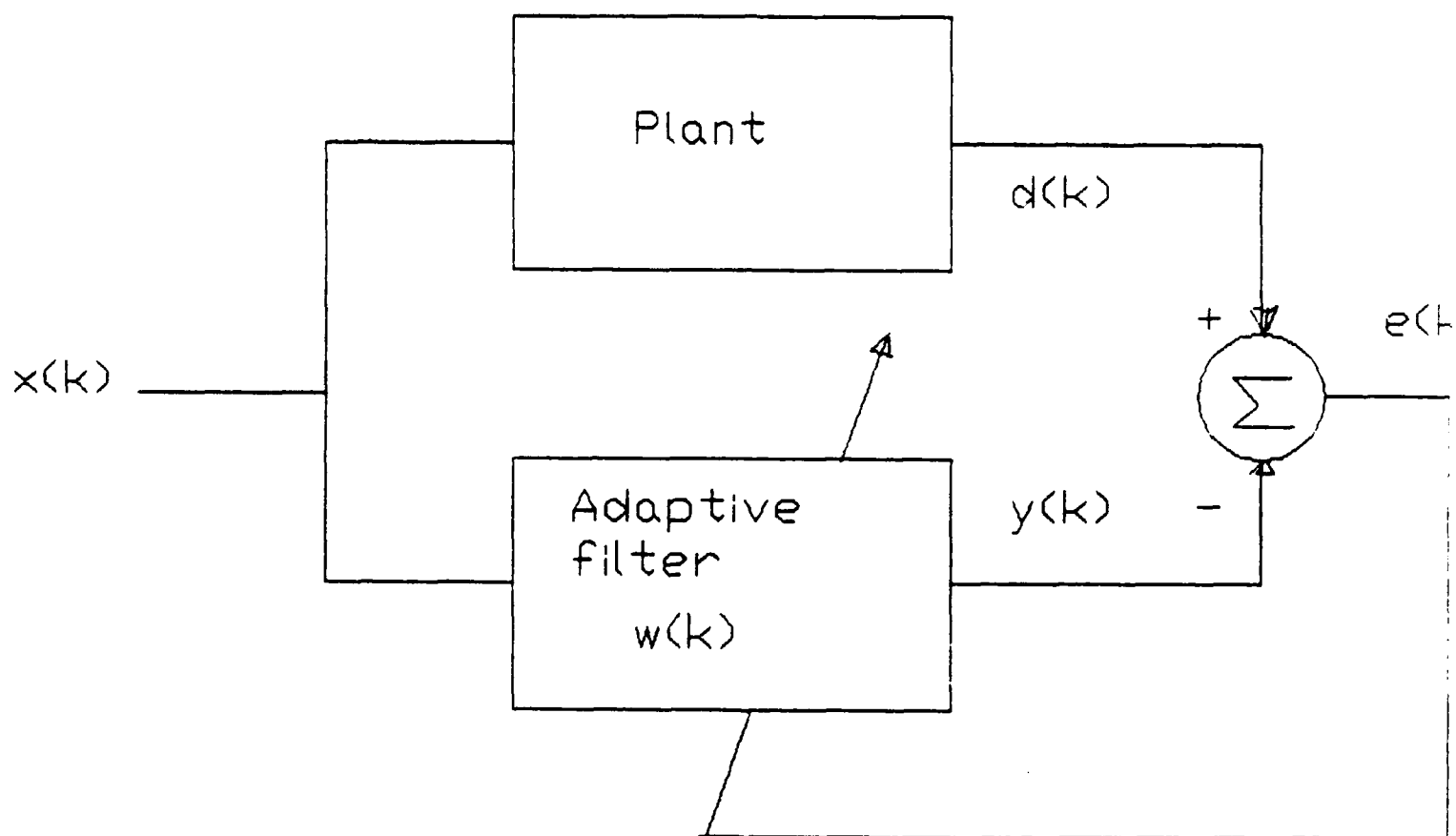


Figure 1: Basic Architecture for Identification

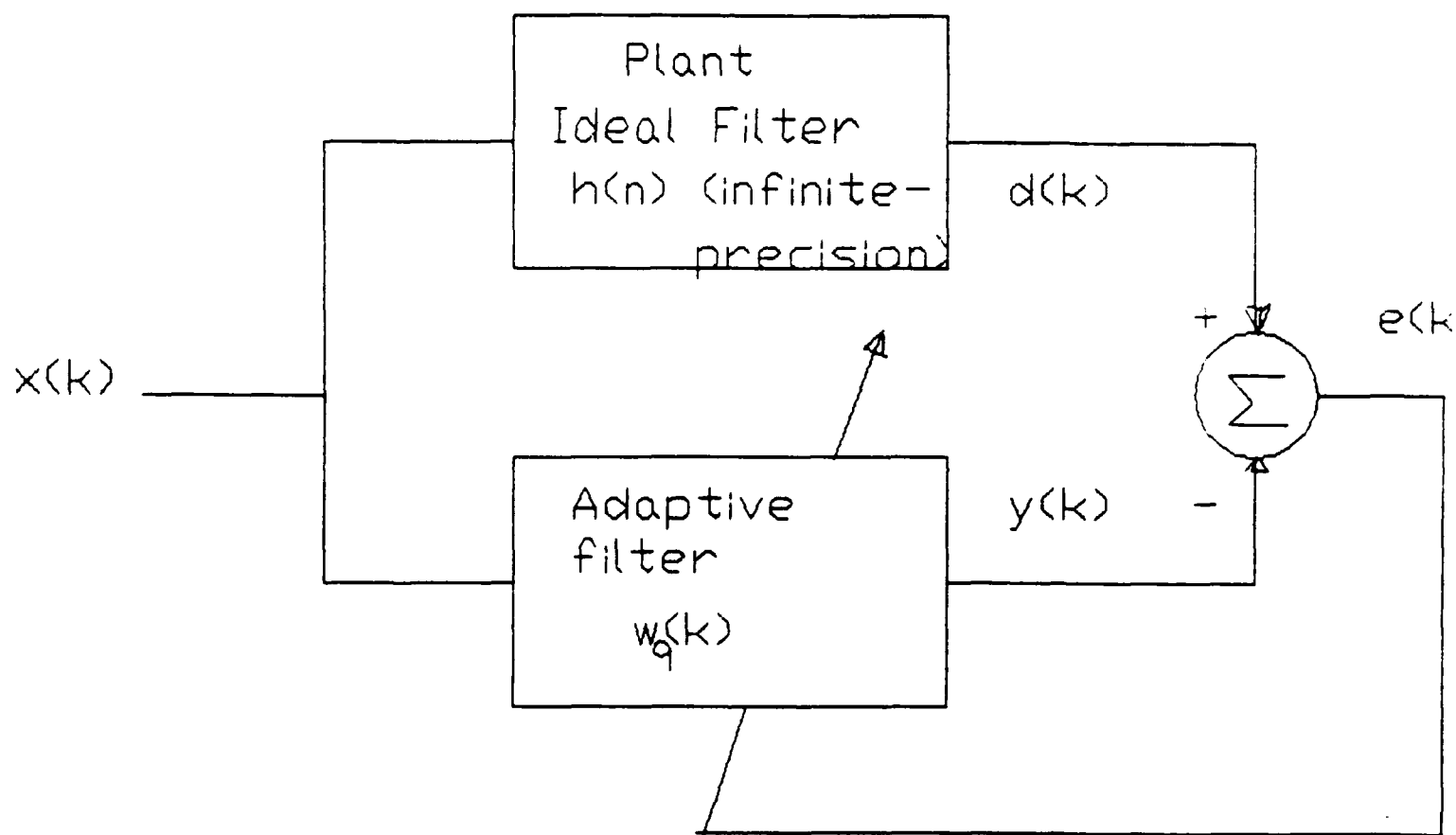


Figure 2: Block Diagram for Finite Wordlength FIR Filter Design

Fig. 3: Magnitude Response of
Length-21 Lowpass Filter
Adaptation Constant = 1.632
Number of Iterations = 500

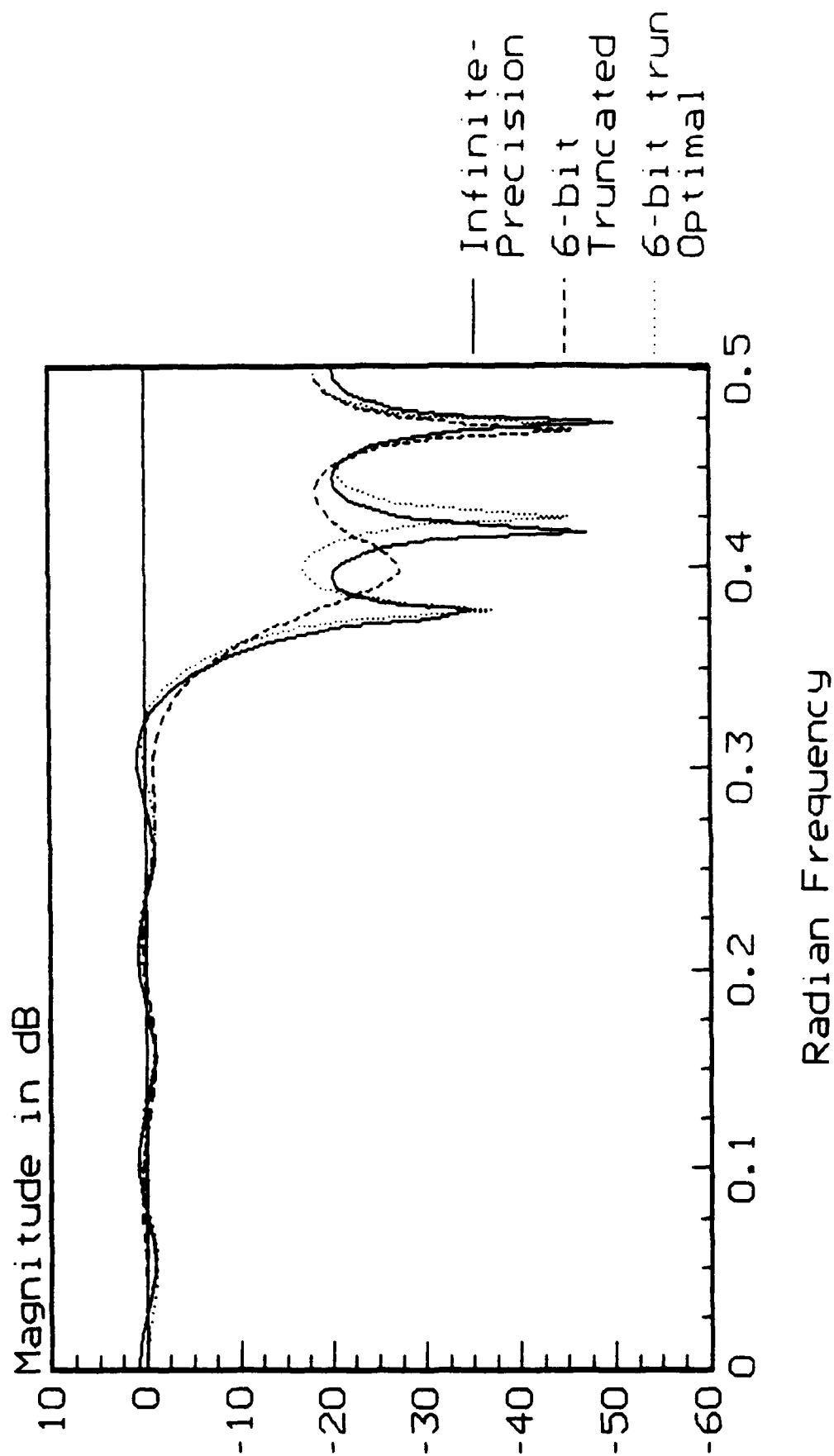


Fig. 4: Error $e(k)$ vs. Iteration No. k

for Length-21 Low-pass Filter

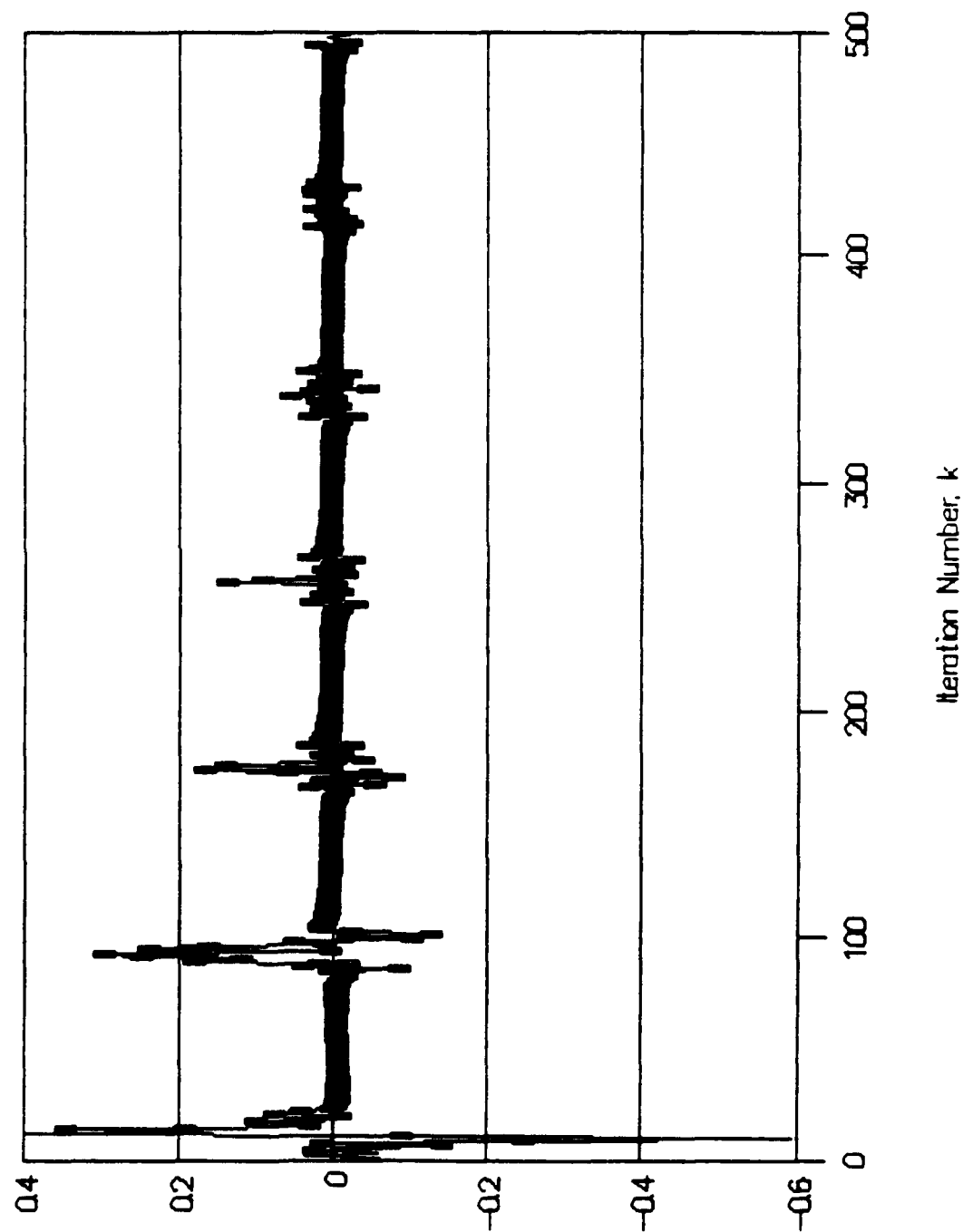


Fig. 5 Magnitude Response of
Length-20 Lowpass Filter
Adaptation Constant = 1.632
Number of Iterations = 500

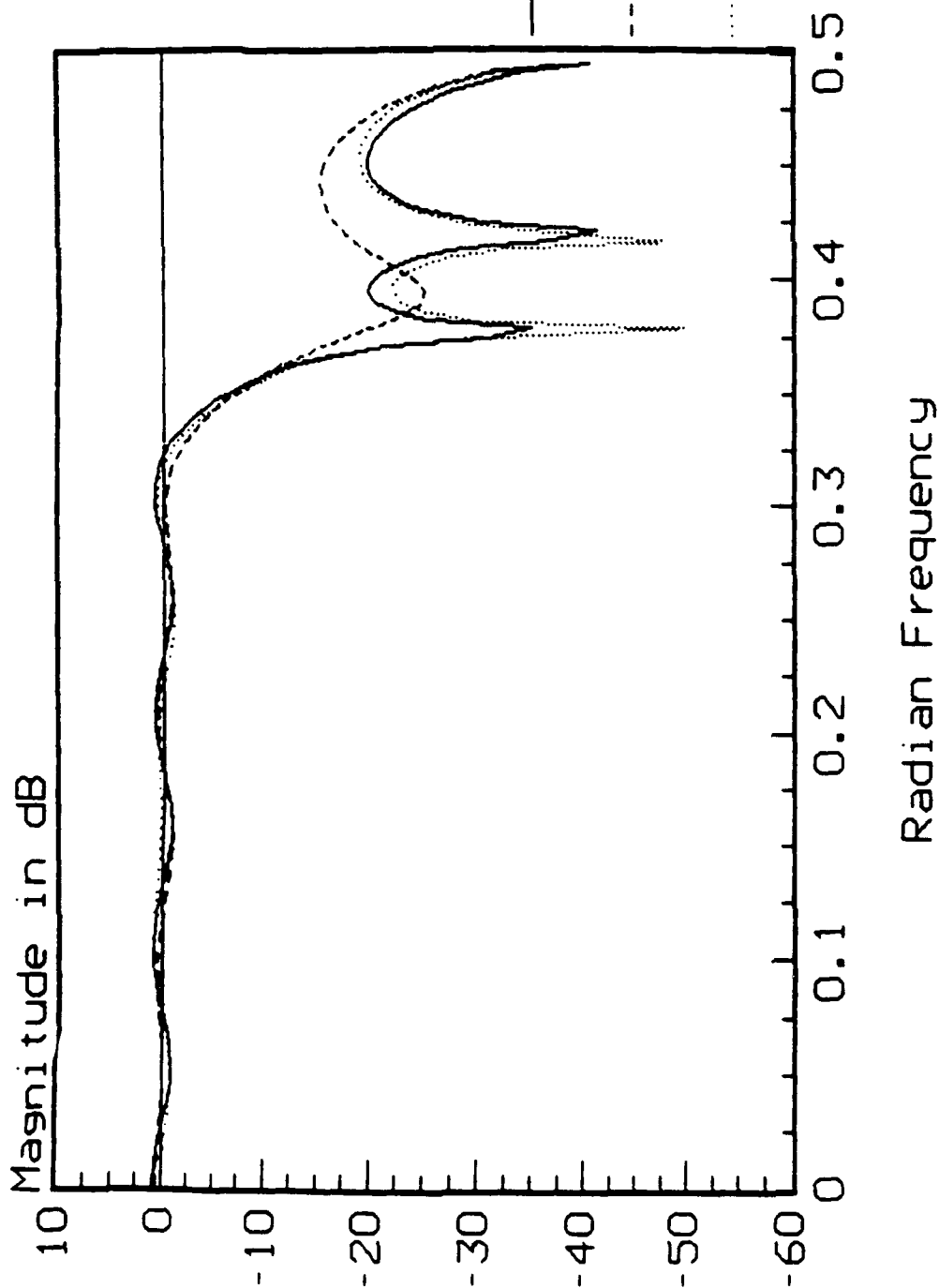


Fig. 6: Error $e(k)$ vs. Iteration No. k
for Length-20 Low-pass Filter

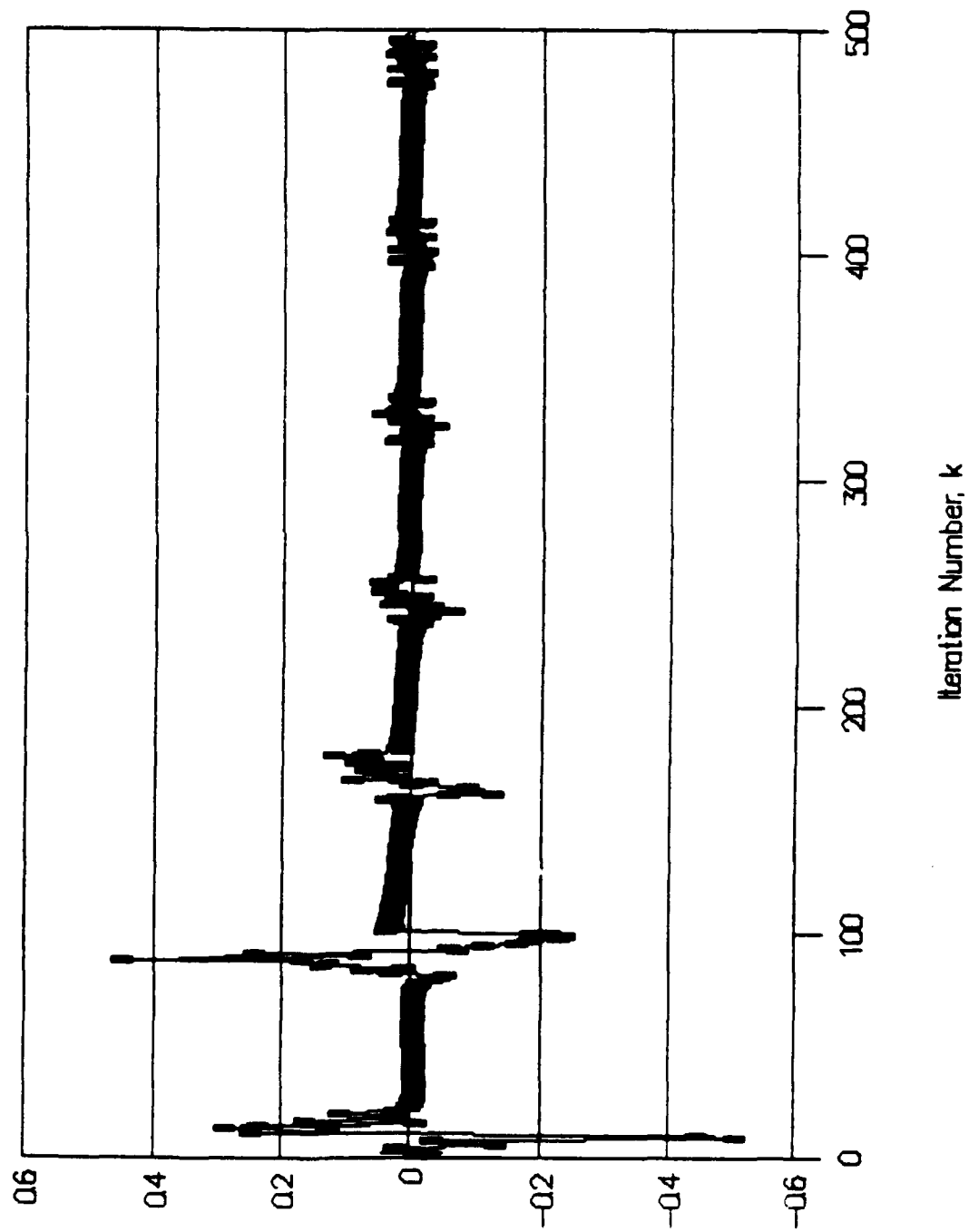


Fig. 7: Magnitude Response
 Length-21 Bandpass Filter
 Adaptation Constant = 1.0
 Number of Iterations = 500

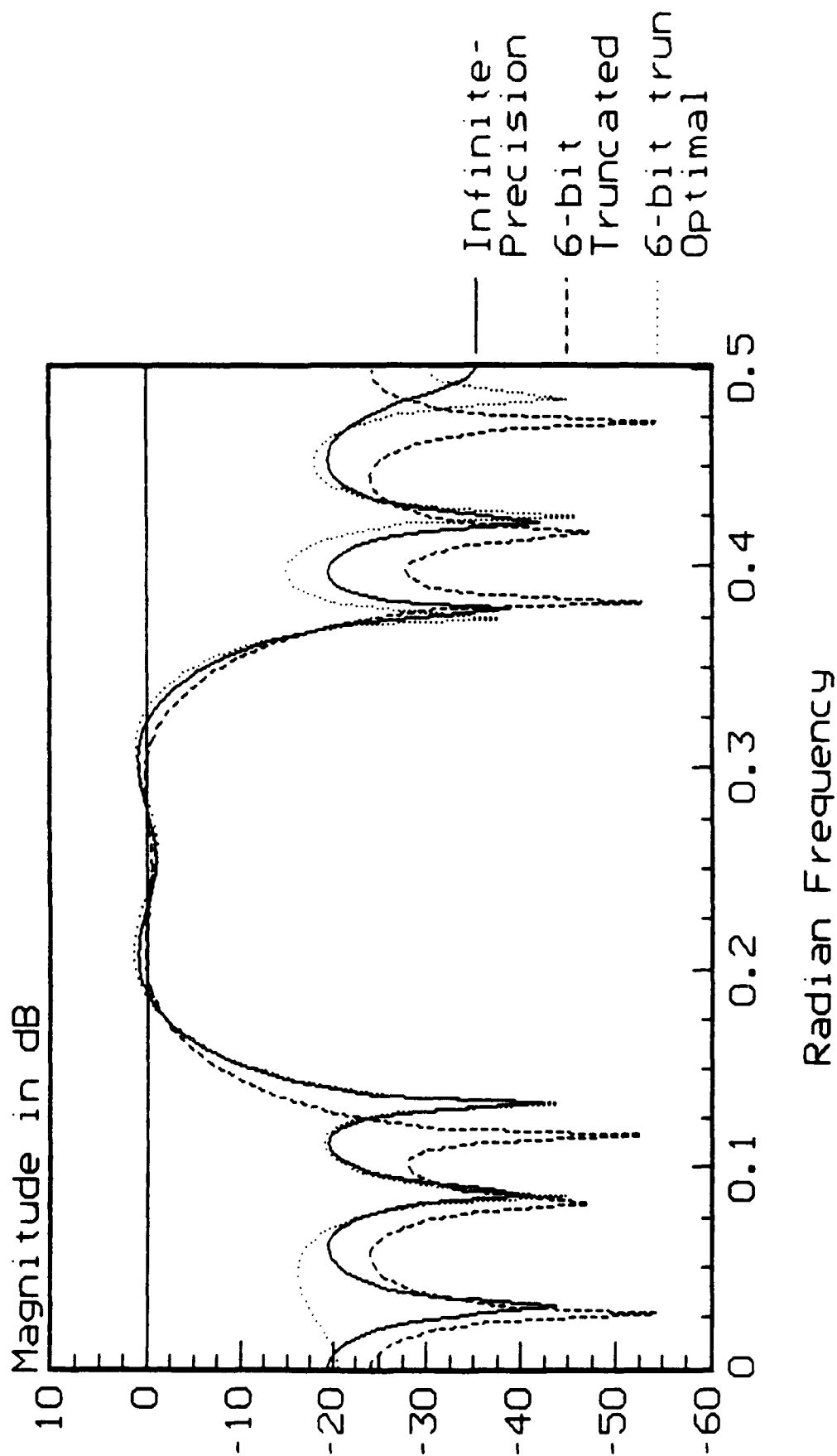


Fig. 8: Error $e(k)$ vs. Iteration No.
for Length-21 Band-pass Filter

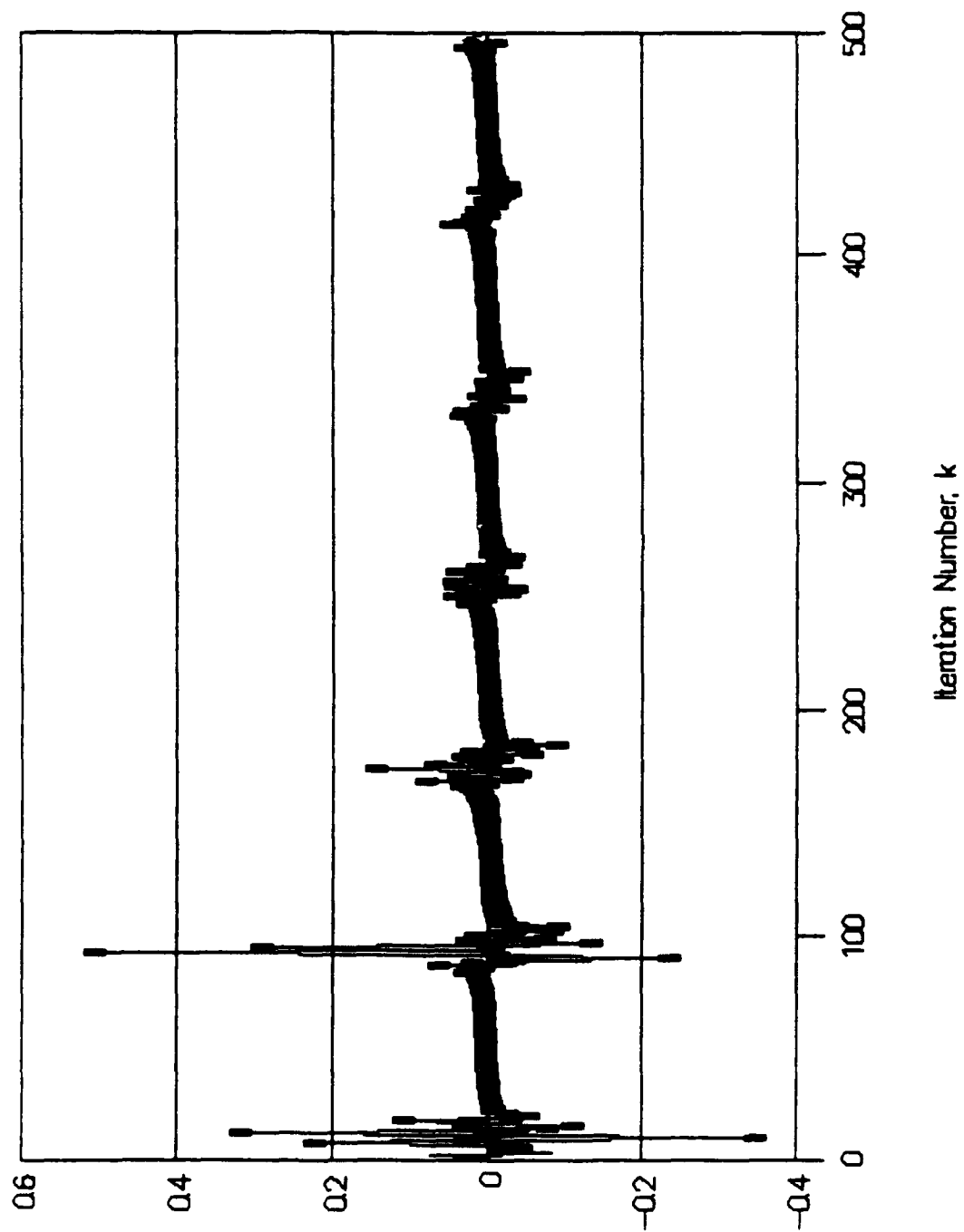


Fig. 9 Magnitude Response of
Length-32 Bandpass Filter
Adaptation Constant = 0.707
Number of Iterations = 1000

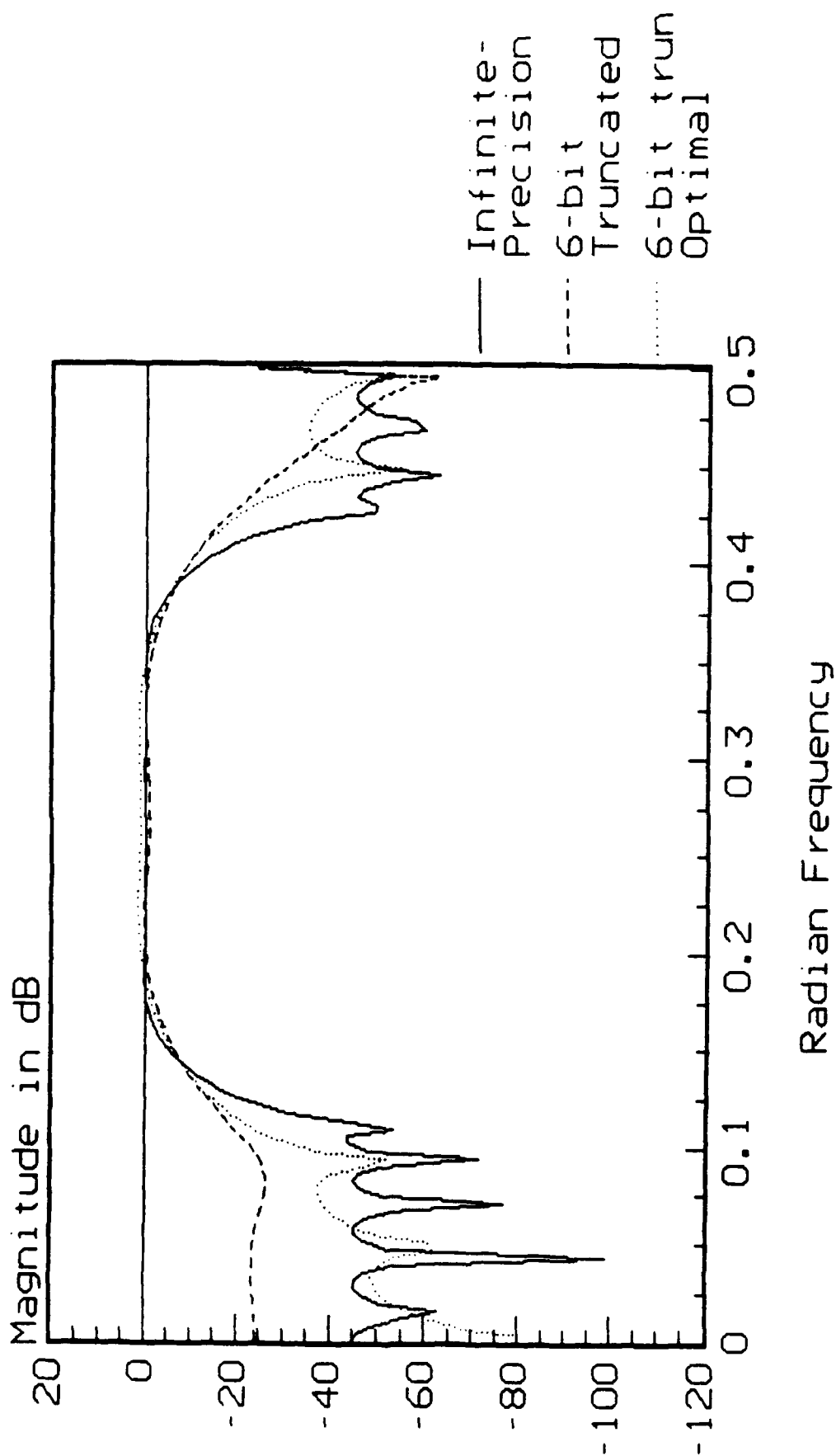


Fig. 10: Error $e(k)$ vs. Iteration No. k

for Length-32 Band-pass Filter

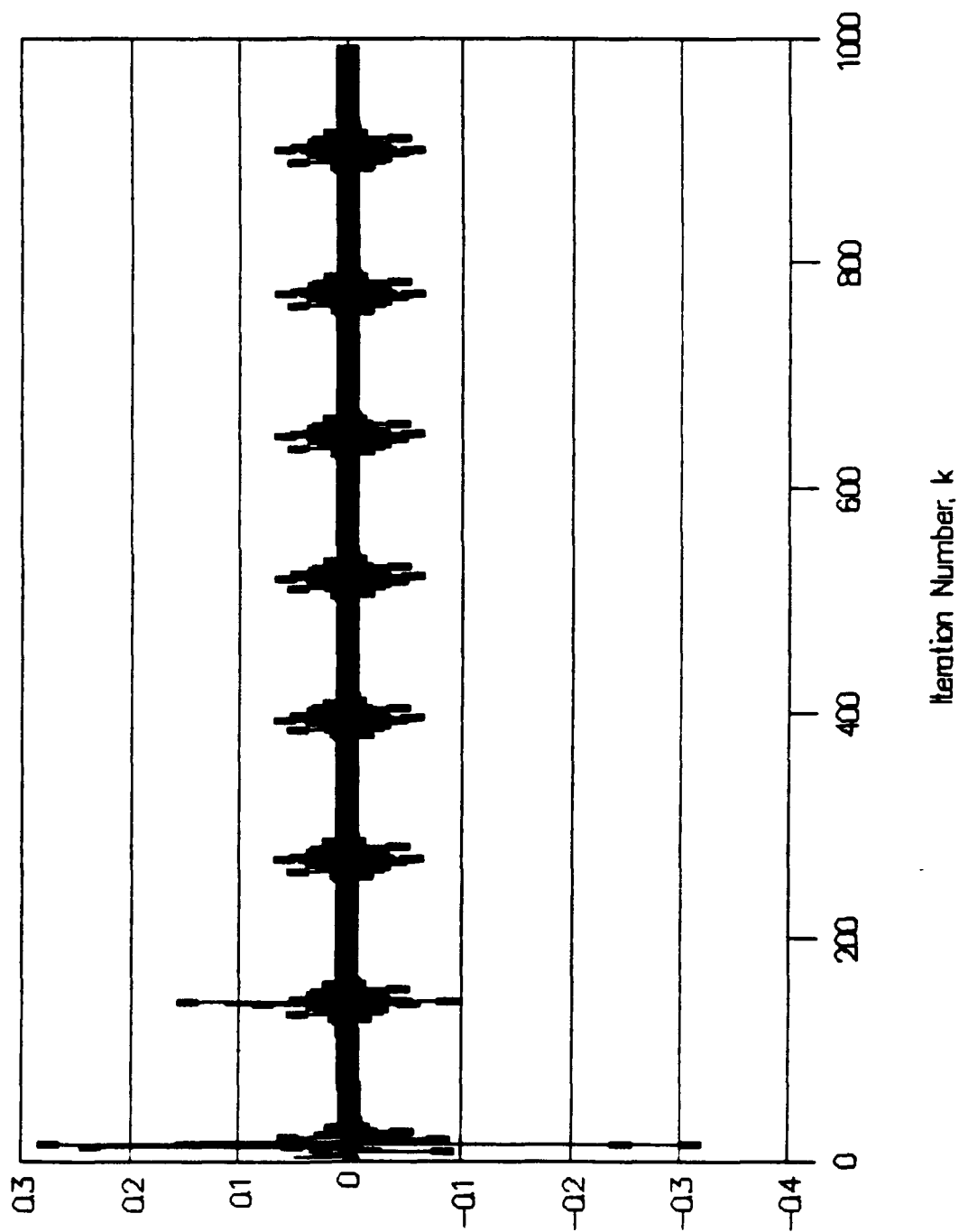


Fig. 11 Magnitude Response
of Length-64 Bandpass Filter
Adaptation Constant = 2.0
Number of Iterations = 1000

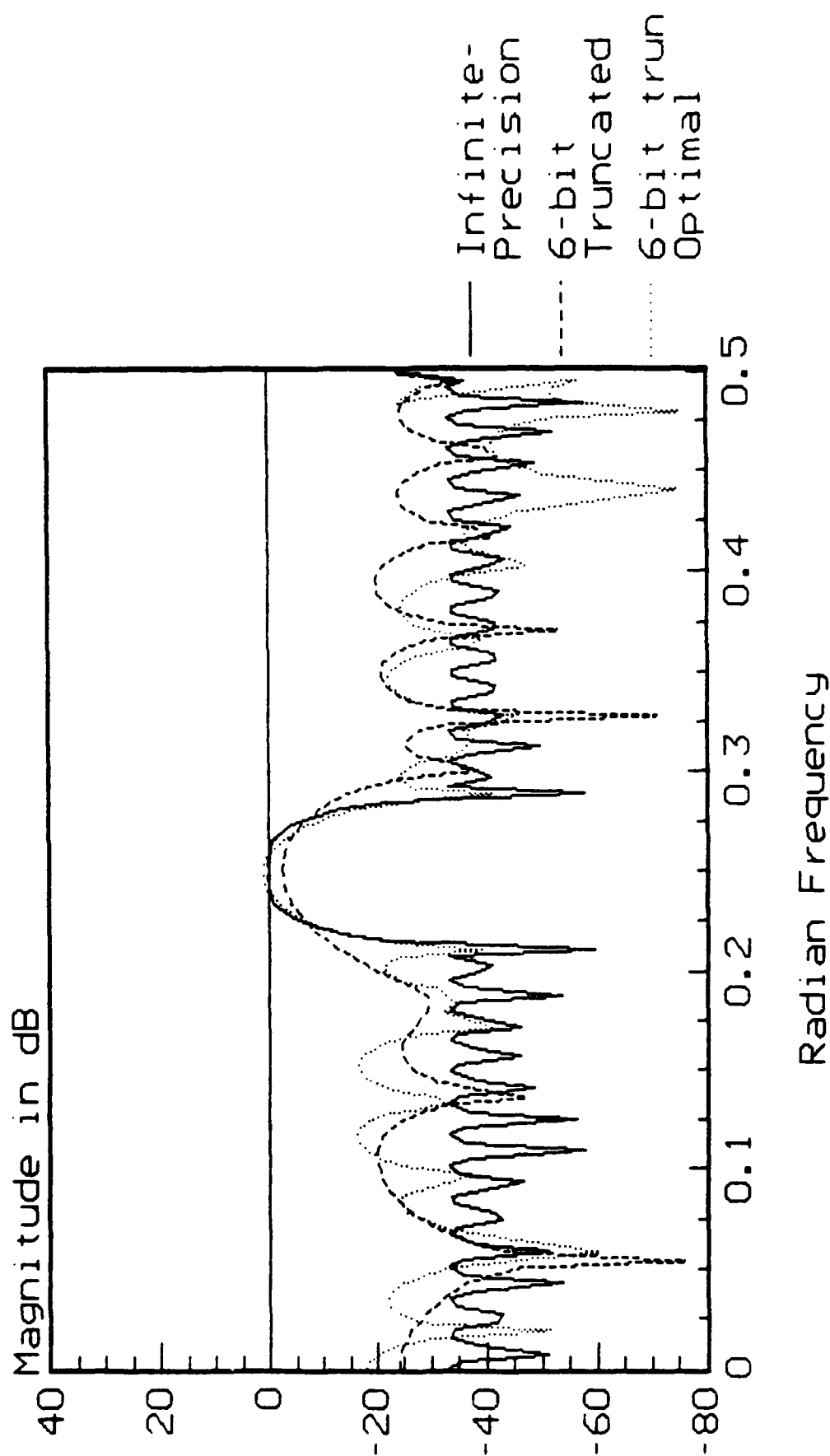


Fig. 12: Error $e(k)$ vs. Iteration No. k
for Length-64 Band-pass Filter

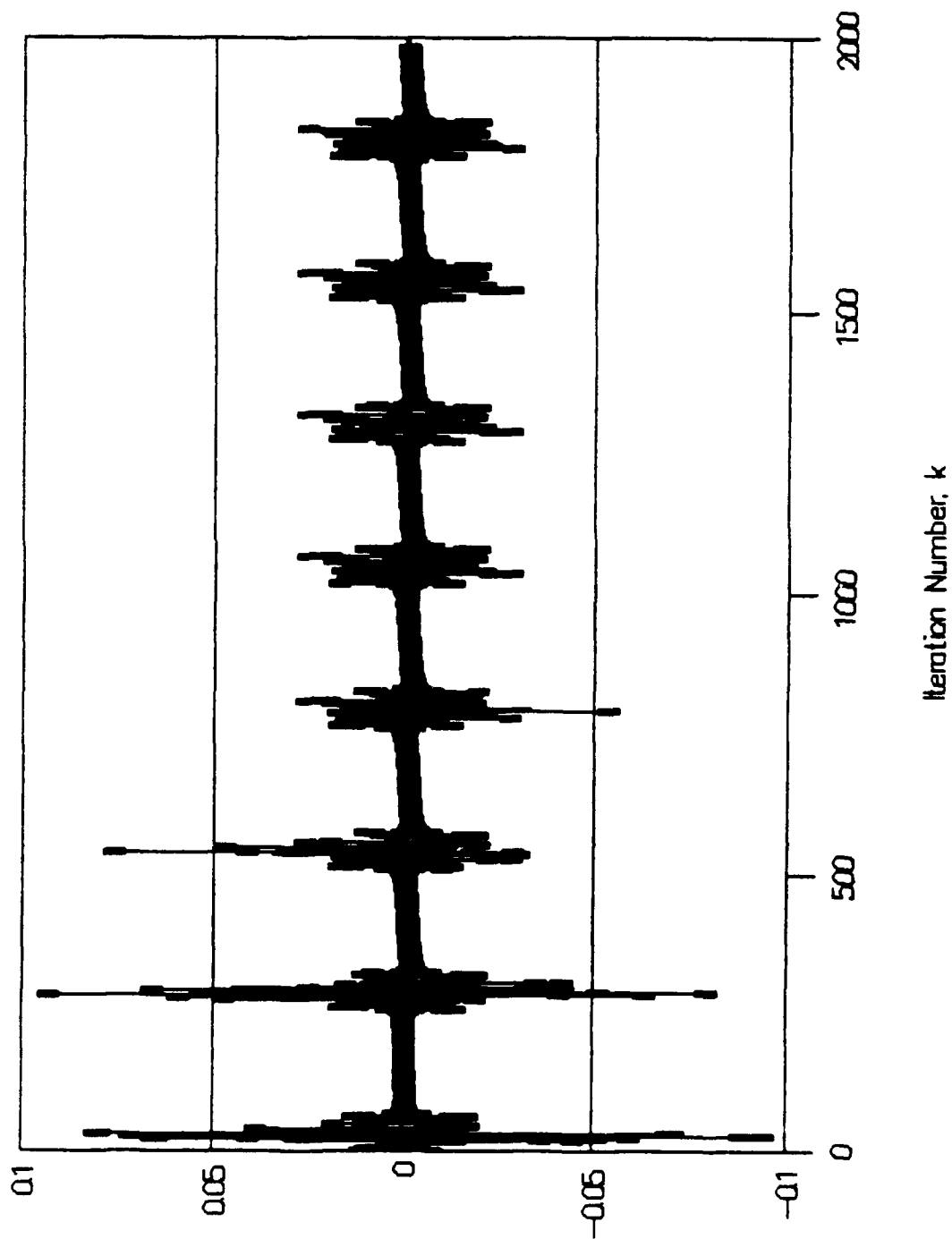


Fig. 13: Magnitude Response of
 Length-31 Bandstop Filter
 Adaptation Constant = 1.0
 Number of Iterations = 1000

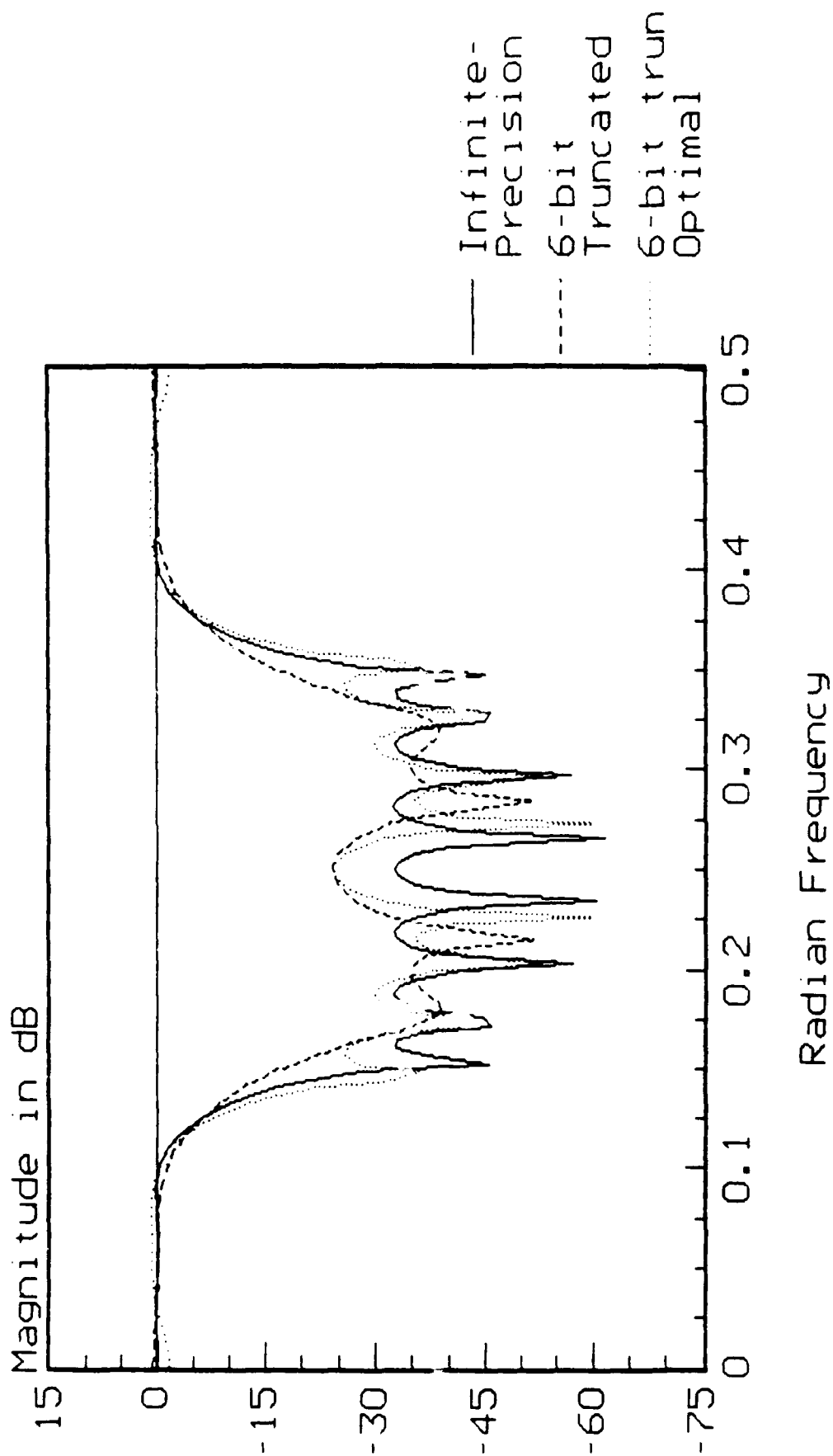


Fig. 14: Error $e(k)$ vs. Iteration No. k
for Length-31 Band-stop Filter

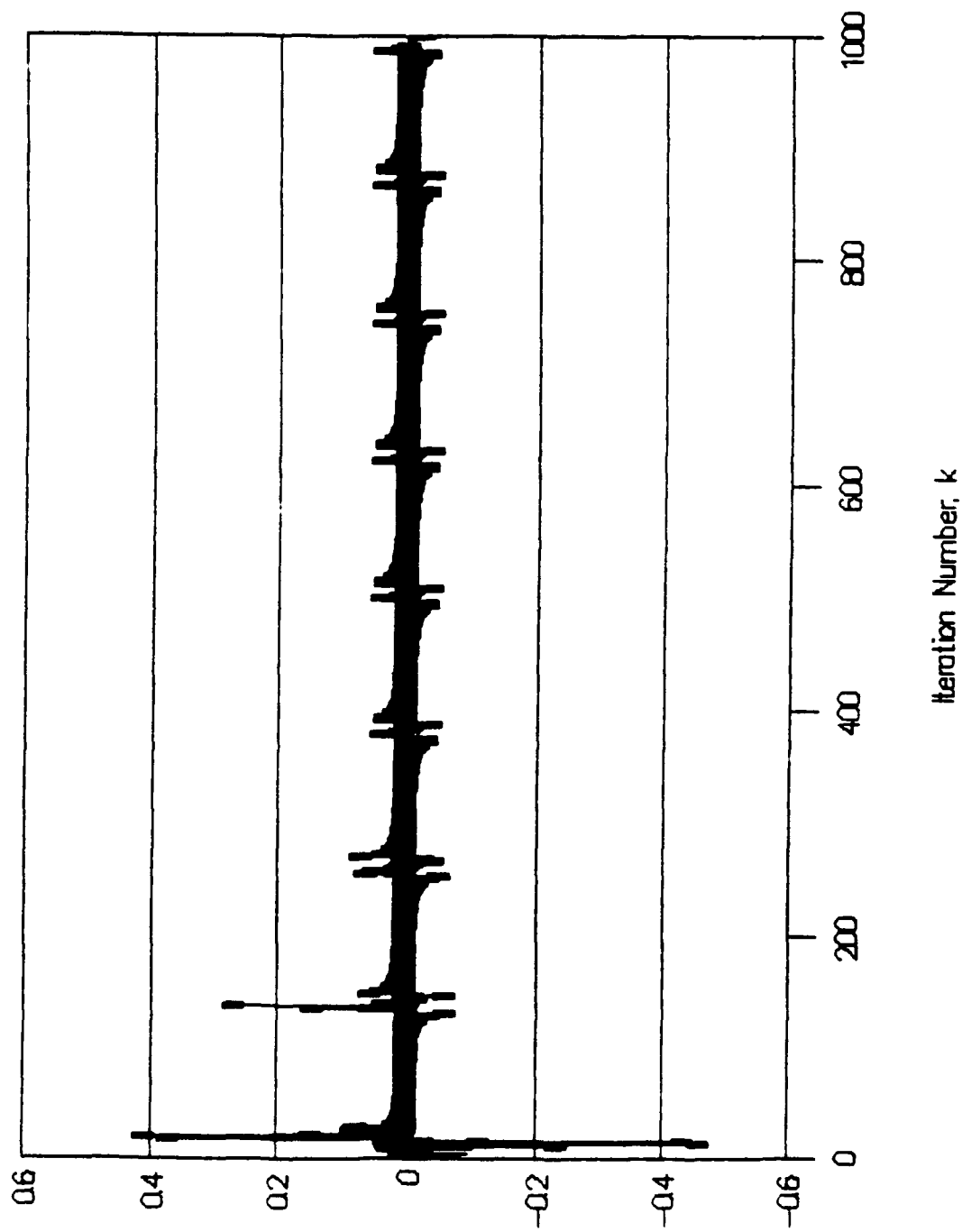


Fig. 15: Magnitude Response of Length-21
Lowpass Filter (Optimal Design w/Defects)
Adaptation Constant = 1.632
Number of Iterations = 500

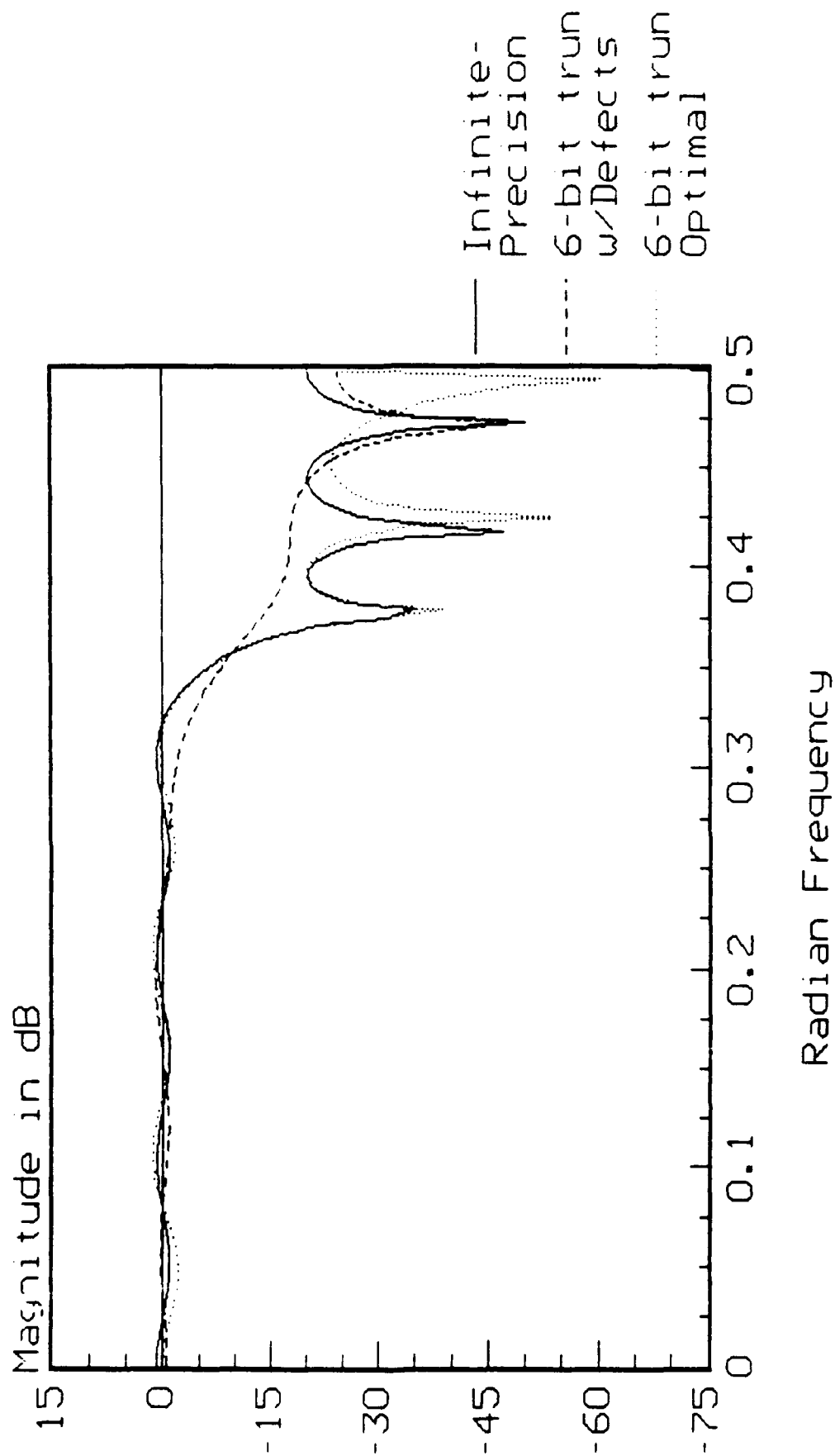


Fig. 16: Error $e(k)$ vs. Iteration No.

for Length-21 Low-pass Filter

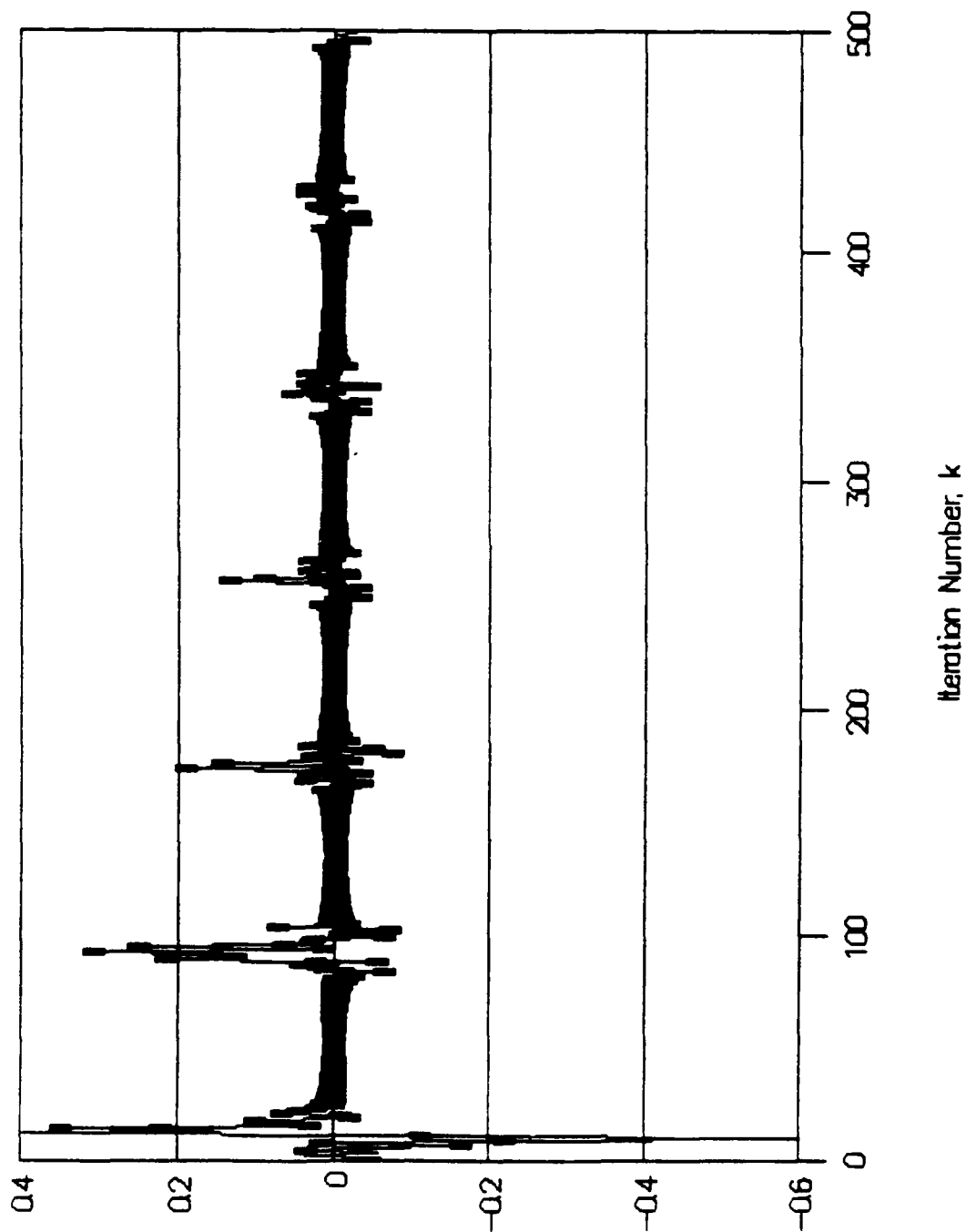


Fig. 17: Magnitude Response of Length-20
Lowpass Filter(Optimal Design w/Defects)
Adaptation Constant = 1.0
Number of Iterations = 500

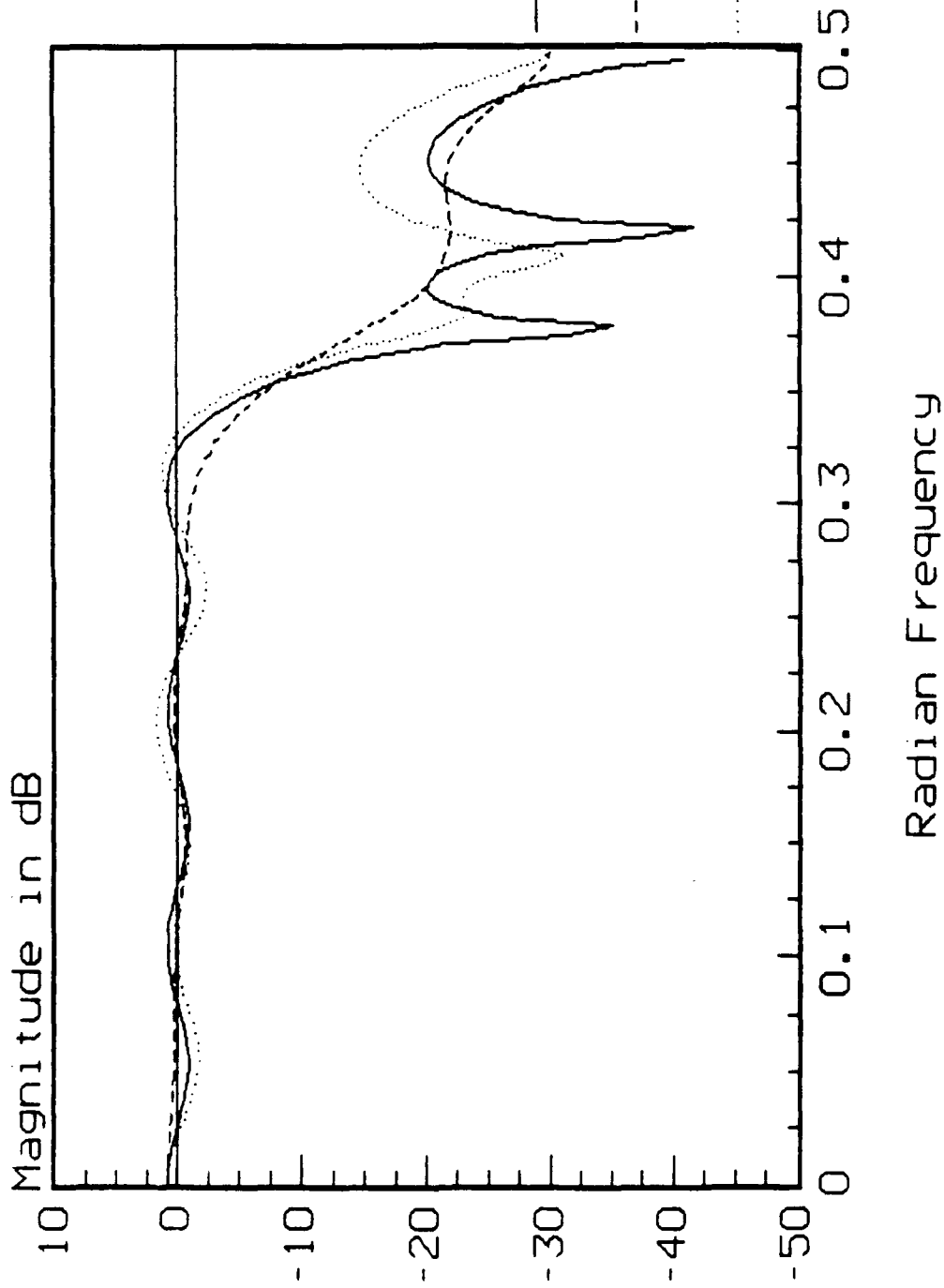


Fig. 18: Error $e(k)$ vs. Iteration No. k
for Length-20 Lowpass Filter

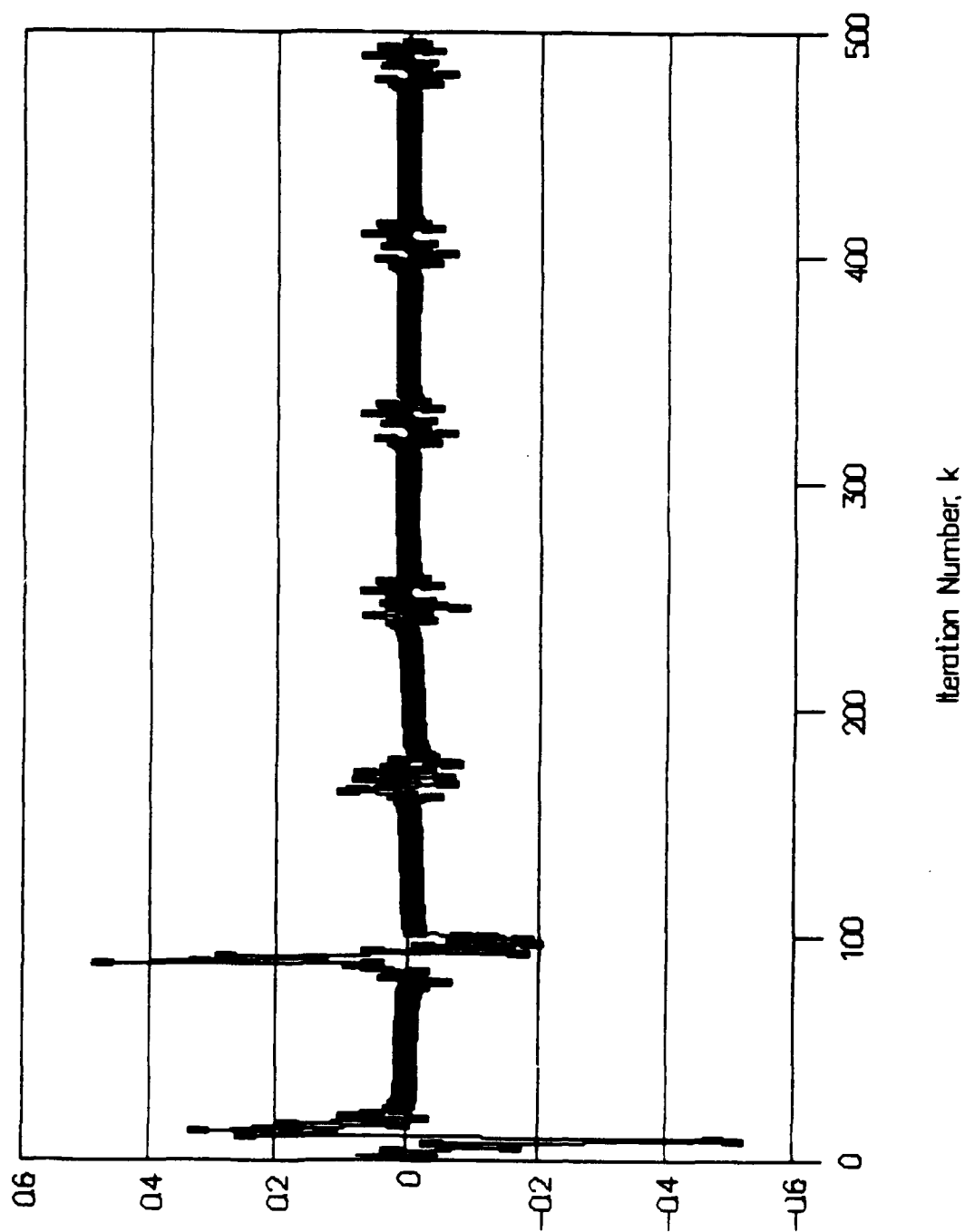


Fig. 19: Magnitude Response of Length-32
 Bandpass Filter<Optimal Design w/Defects
 Adaptation Constant = 0.707
 Number of Iterations = 1000

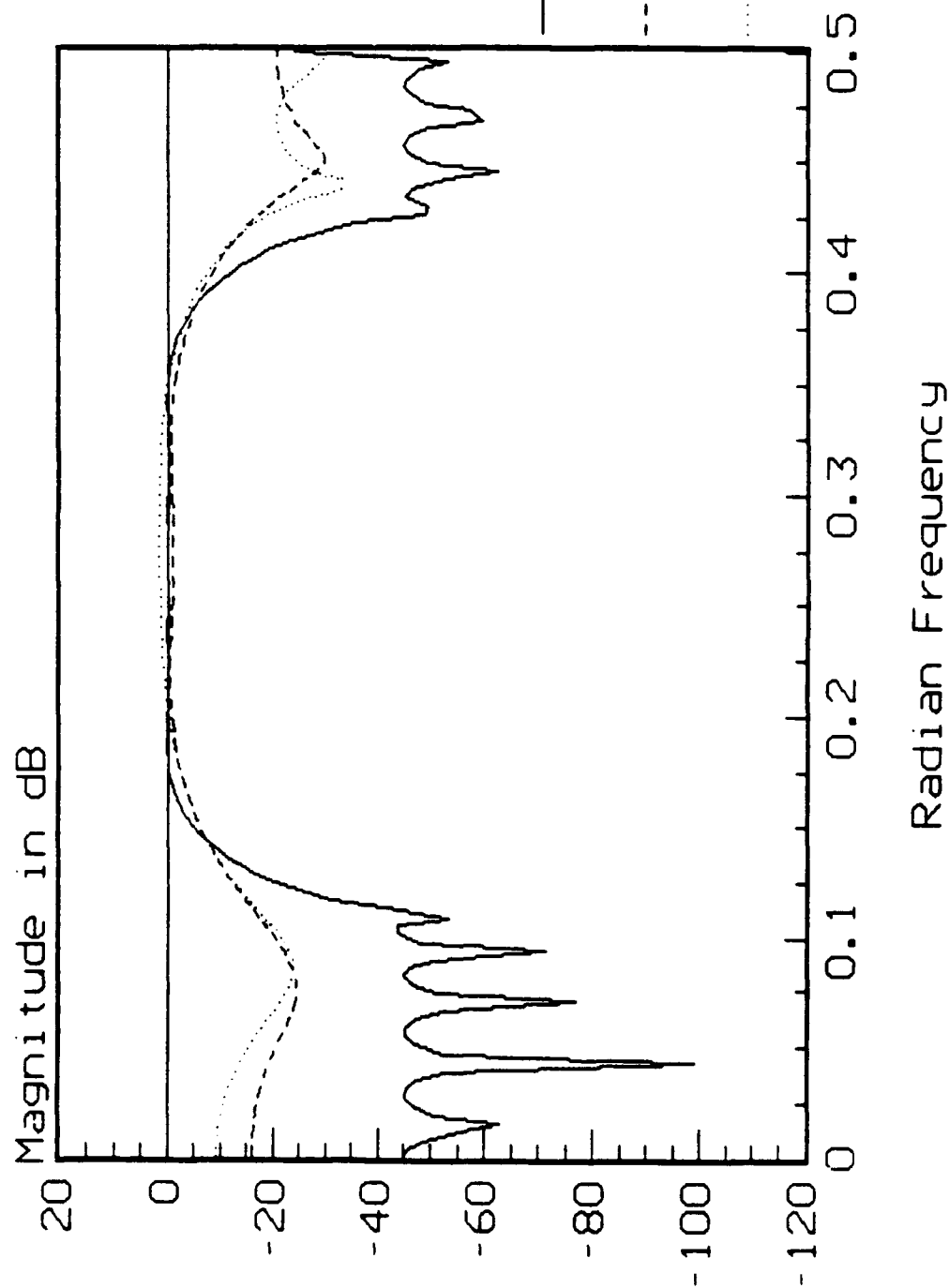


Fig. 20: Error $e(k)$ vs. Iteration No. k

for Length-32 Band-pass Filter

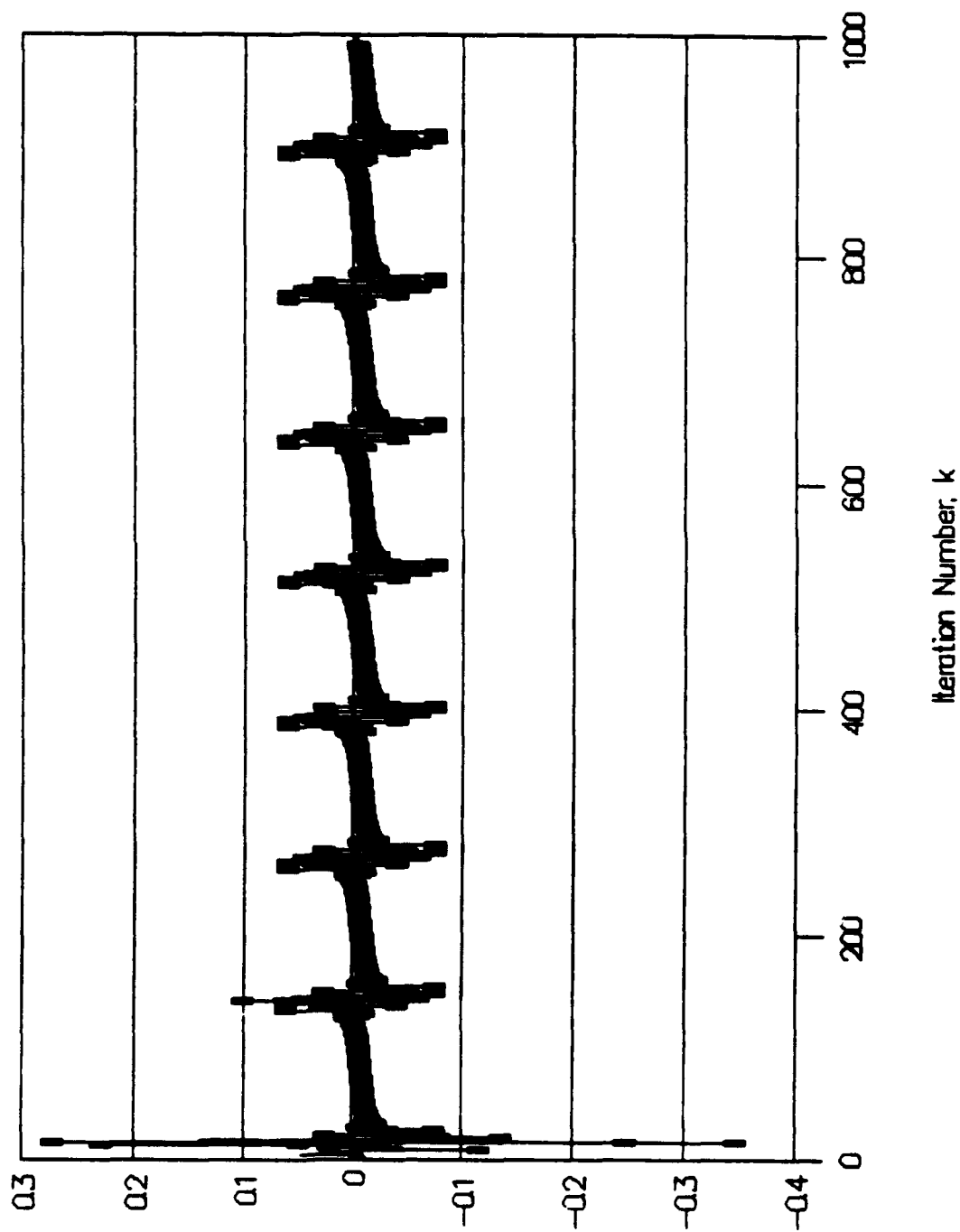


Fig. 21: Magnitude Response of Length-64
 Bandpass Filter<Optimal Design w/Defects
 Adaptation Constant = 2.0
 Number of Iterations = 2000

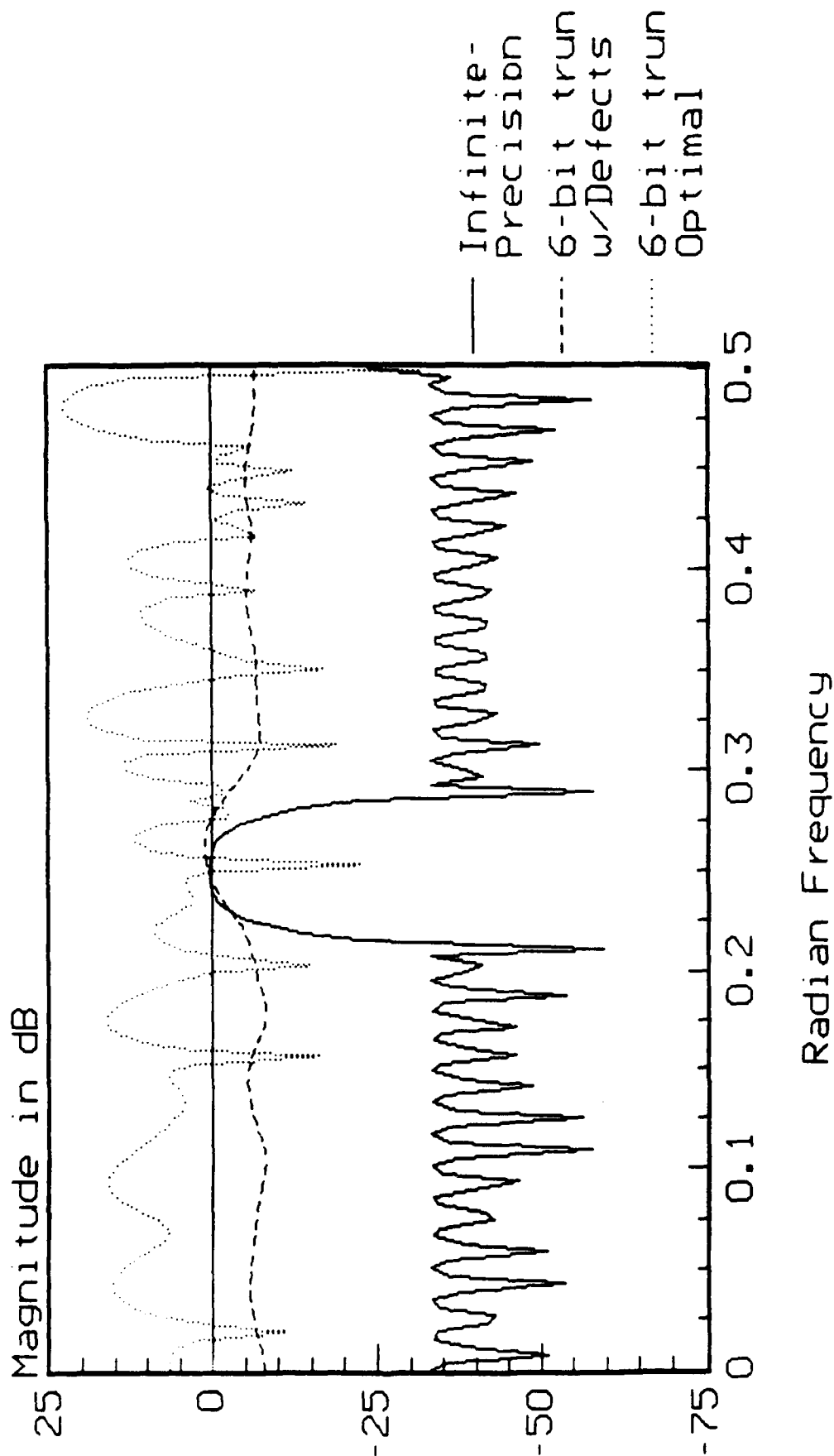


Fig. 22: Error $e(k)$ vs. Iteration No. k

for Length-64 Band-pass Filter

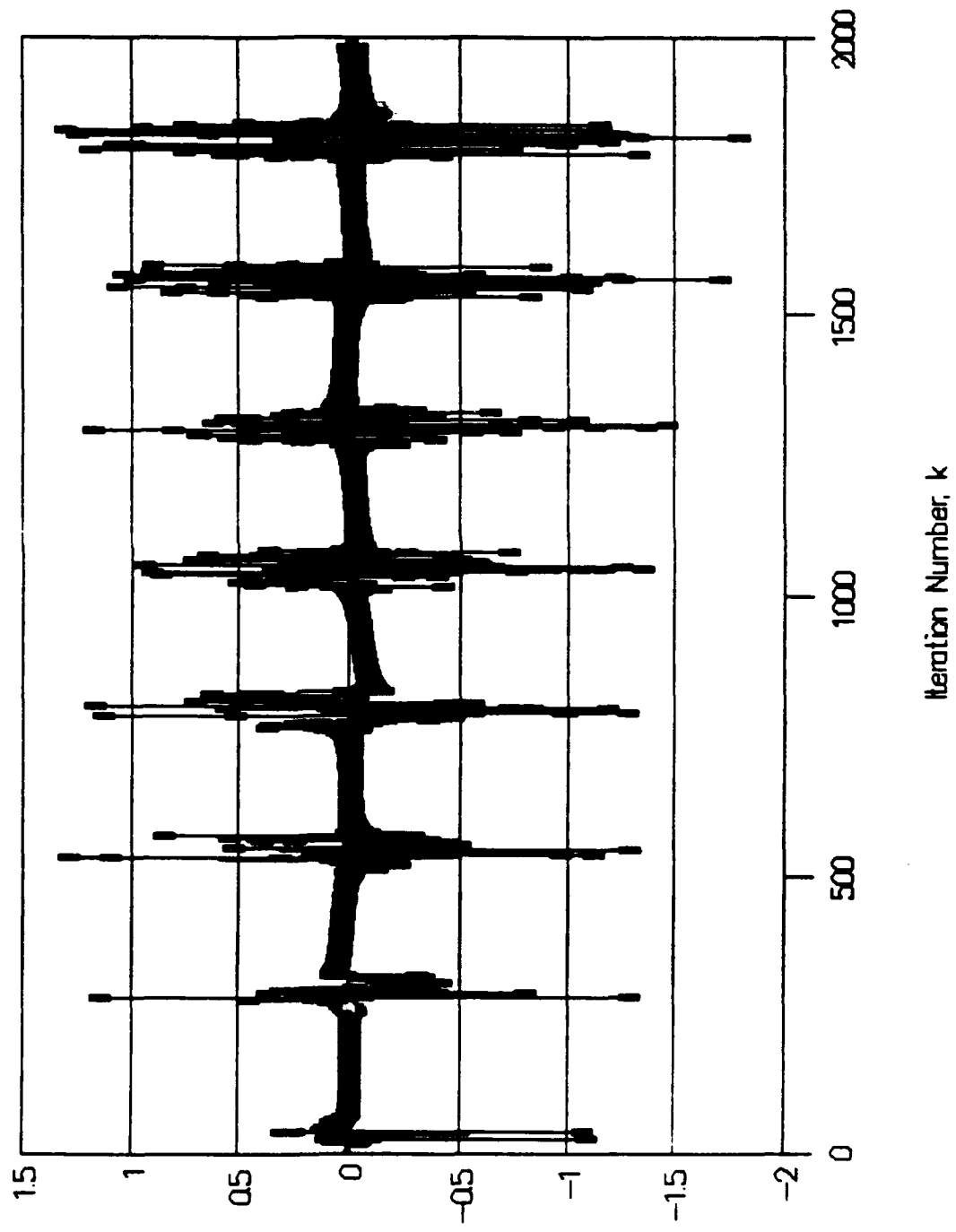


Fig. 23: Magnitude Response of Length-31
 Bandstop Filter<Optimal Design w/Defects
 Adaptation Constant = 1.0
 Number of Iterations = 1000

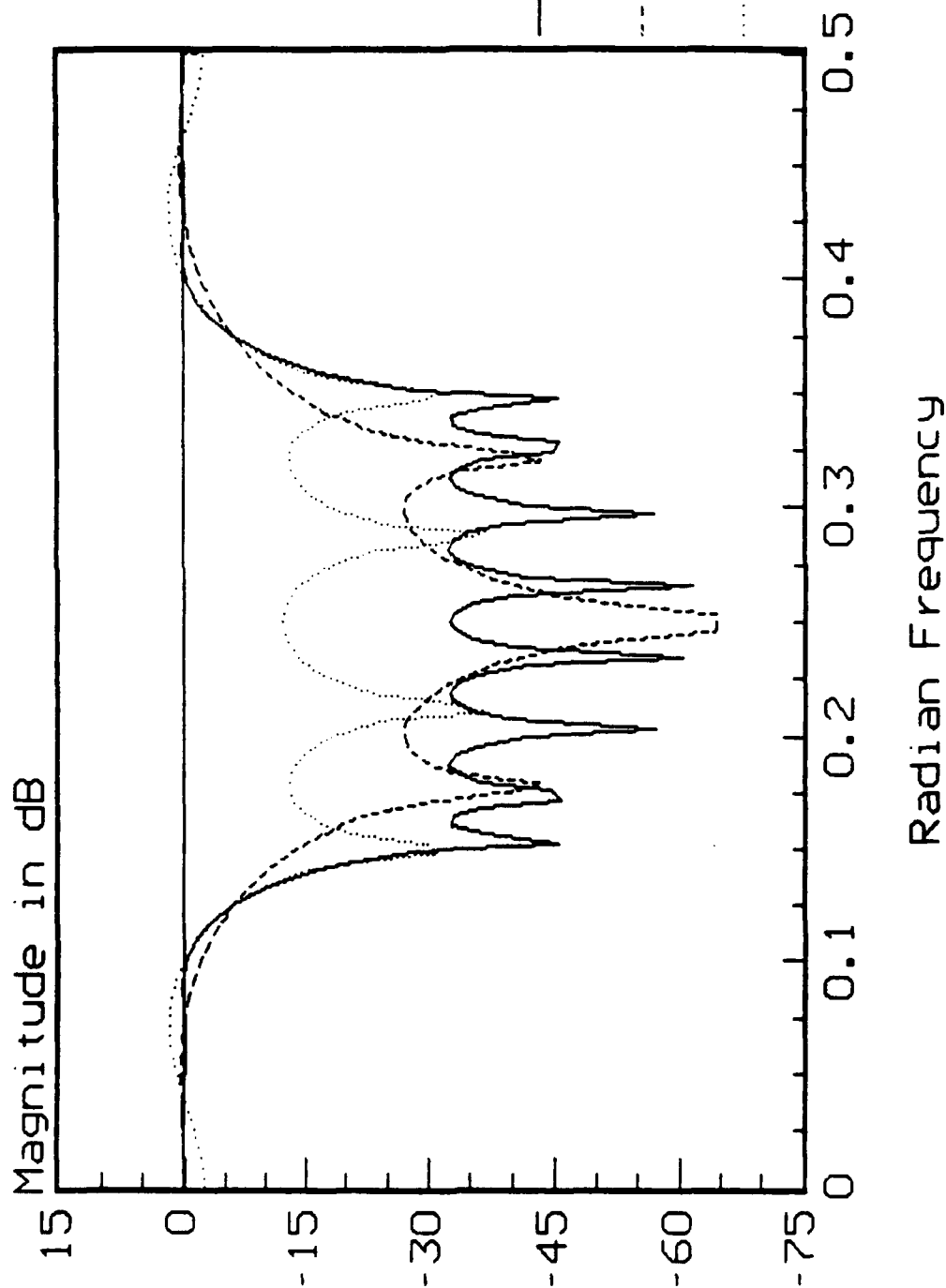
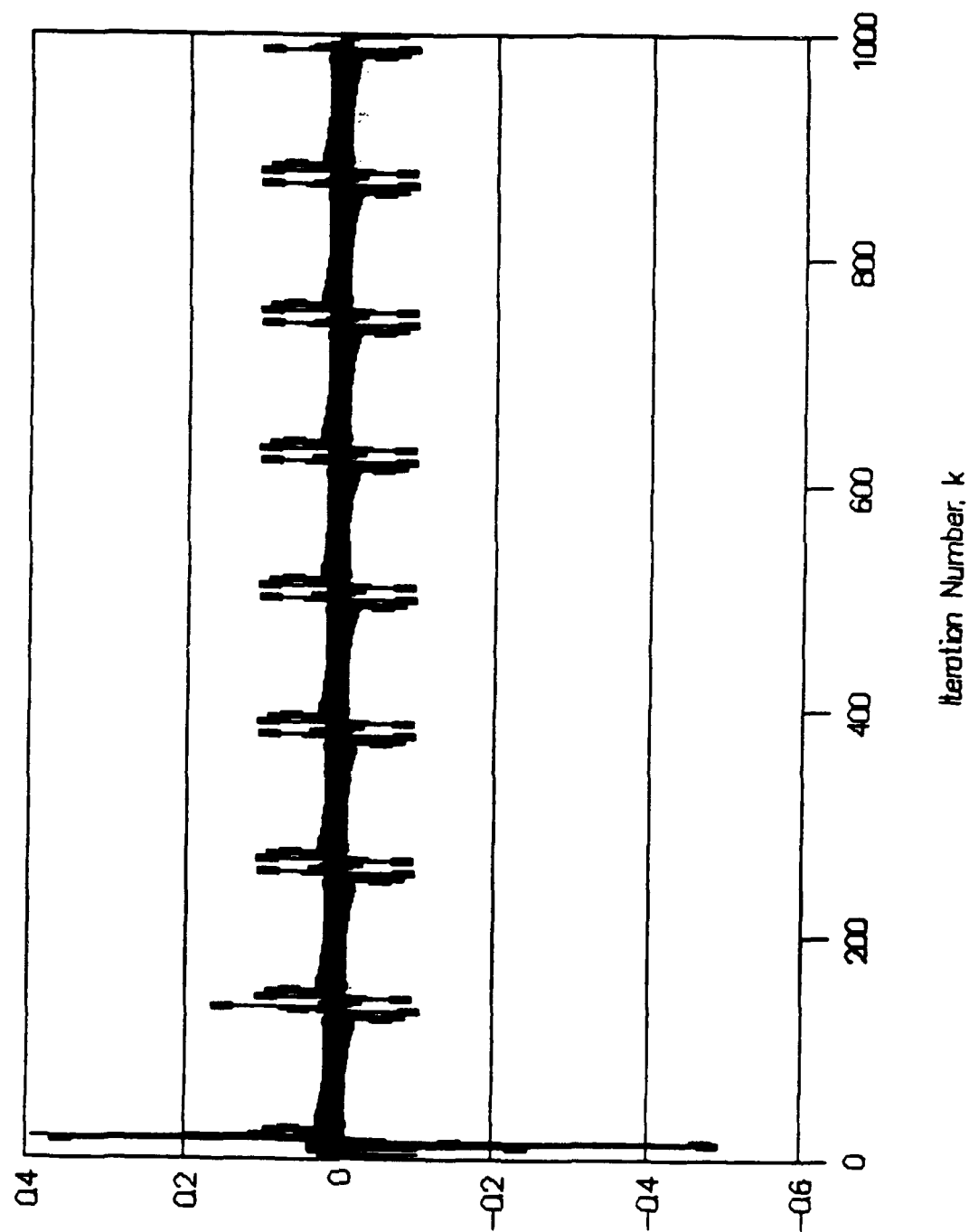


Fig. 24: Error $e(k)$ vs. Iteration No. k
for Length-31 Band-stop Filter



**Automatic Adaptive Remeshing for Finite
Element Reliability Assessment of
Electronic Devices**

Ian R. Grosse

**Final Report
to
Computer Aided Engineering Systems Group
Rome Air Development Center
on
Universal Energy Systems Grant No. S 210 10 MG 129
(UMASS Account No. 5-22966)**

February 1991

**Department of Mechanical Engineering
University of Massachusetts
Amherst, MA 01003**

Abstract

Reliability prediction of microelectronic devices using the Finite Element Method requires accurate estimation and control of the inherent finite element discretization error. This report details formal mathematical approaches by which the finite element discretization error may be automatically estimated in a computational environment, and the results of this assessment used to automatically control this inherent numerical error. To this end, two new *posteriori* finite element error estimators and automatic adaptive mesh generation algorithms were developed and implemented into FORTRAN codes for 2-D automatic adaptive mesh refinement. Both methods assess the discretization error element by element by comparing the discontinuous finite element scalar function with an improved piecewise continuous (C^0 continuous) scalar function obtained through postprocessing of the finite element results. The two error estimators are distinguished from each another by the manner in which the C^0 continuous scalar function is obtained. Results are presented for 2-D elasticity in which the effective stress (*von Mises*) is taken as the scalar function field and the corresponding L_2 norm has a distortional energy interpretation. Finally, the concept of adaptive accuracy is introduced in a h -based adaptivity algorithm. Examples are presented which demonstrate the effectiveness of the error estimation and control algorithms as implemented in a computational system.

1 Introduction

The finite element method is the most widely used numerical method for obtaining approximate solutions to partial or ordinary differential equations which govern complex engineering problems. The Computer Aided Systems Engineering Branch (RBES) at Rome Air Development Center has successfully employed the finite element method for reliability assessment of microelectronic components ([1],[2]). Detailed finite element modeling and analysis of semiconductor chips, leads, weld joints, etc., has yielded accurate predictions of the location and magnitude of critical stresses and temperatures in the microelectronic component ([3]-[14]). Based on these results, failure modes can be predicted and the mechanical reliability of the product assessed. While there is no substitute for statistical and empirical reliability prediction methods, such as those found in Military Standard MIL-HDBK-217E, finite element analysis of microelectronic components can supplement these reliability prediction methods by offering a deeper understanding of the physics of design-related failure modes. With this insight the engineer can then make design changes to improve the reliability of the device.

The finite element reliability prediction method is particularly useful as a reliability design tool for proposed new technology or custom devices where one does not have the benefit of extensive (or any) experimental data. Often, these devices are quite costly and time consuming to develop and manufacture. It is critical, therefore, that design-related reliability problems be revealed and corrected early in the prototype design phase, before expensive tooling and processing costs have been incurred. Finite element analysis of the device can give the designer and engineer the knowledge needed to make the best possible design decisions in the absence of hard experimental data.

However, reliability assessment via the finite element method has been greatly limited by the sub-

stantial amount of labor-intensive work and finite element modeling expertise required for accurate finite element analysis. Typically, it may take an engineer several weeks to build an initial detailed finite element model of the microelectronic device on the computer. Moreover, the numerical results will often indicate critical regions of the device which must be remodeled in greater detail to obtain an accurate solution for these regions. Accordingly, finite element reliability assessment (FERA) is inherently an iterative, and currently a time-consuming, process. A flow diagram of this process as implemented today is shown in Figure 1. Several iterations are usually required before accurate results can be obtained. The shaded box indicates the various tasks involving significant human interaction. Of these, the task of building geometric models and generating valid finite element meshes are typically the most labor-intensive and time-consuming, although assessing the mechanical reliability of the microelectronic device using the finite element results can also be a cumbersome process.

The above discussion leads to the following observations. First, reliability predictions of microelectronic devices by the finite element method are meaningless *unless the finite element analysis itself is reliable*. Second, reliable finite element analysis is inherently an iterative process. In the absence of empirical data, a single analysis is clearly insufficient for guaranteeing sufficiently accurate results. Lastly, the utility of the finite element method as a design and reliability prediction tool is greatly limited by the inefficient, labor-intensive iterative process of analysis, refinement, reanalysis, etc.

Obviously, there is a great need to automate this iterative process. Fortunately, research over the past ten years has begun to address this problem through the development of various formal algorithms for estimating the finite element discretization error and for automatically refining the mesh based on these error estimates. Methods which estimate the finite element discretization

error based on the analysis results are called *A-posteriori* Error Estimators. The process of refining the mesh either locally or globally based on *a-posteriori* error estimates is called Adaptive Mesh Refinement. This report details the research involved in a one year project on automating the process of converging on a sufficiently accurate finite element solution.

It should be noted that the process of actual mesh generation has been greatly facilitated by the emergence of powerful automatic mesh generators which can now generate nonuniform meshes based on specified mesh densities at specific points. The advent of these automatic mesh generators has made automatic adaptive mesh refinement possible for finite element codes.

2 Objective

The objective of this research project is to significantly improve and increase the accuracy and efficiency of finite-element based reliability assessment of microelectronic devices. This can be accomplished by the development of formal error estimation and adaptive remeshing algorithms, the implementation these algorithms into computer codes, and the interfacing of these codes with finite element mesh generation codes, such as FASTQ, and finite element analysis codes, such as NISA2, as shown in Fig. 3. In this manner much of the time-consuming manual effort required to converge on an accurate finite element solution will be eliminated. Furthermore, the computational system will provide the reliability engineer with important estimates of the accuracy of the numerical simulation. This will result in greater productivity and improved finite element reliability assessment of microelectronic devices.

It should be noted that the objective of this research is consistent with a longterm objective of

automated finite element reliability assessment. The proposed research is the first step needed to achieve this longterm objective.

3 Related Work

In this section a brief sampling of related work in *a-posteriori* error estimators and adaptive mesh refinement is presented. For a more comprehensive review the reader is referred to Shephard [15] and Babuska et al. [16].

Turcke and McNeice [17] were one of the first researchers to study the problem of assessing the discretization error. The authors established some simple informal guidelines for using the strain energy density function as a measure of the discretization error. In 1977 Melosh and Marcal [18] proposed the "specific energy difference" (SED) method for assessing the discretization error. If the solution has converged, then the strain energy differential between each element with a reduced degree of freedom set (i.e. from the previous mesh) and the corresponding "same" element with a higher degree of freedom set (i.e. from the current mesh) must be zero. Thus, this strain energy differential is a p -based measure of the discretization error. It can also be viewed as an h -based error analysis method if the corresponding "same" element actually represents a "set" of elements obtained by subdivision of the element from the previous mesh. To avoid two separate analysis, Melosh and Marcal argued that this strain energy differential between elements of different meshes can be approximated from a single solution as the difference between the specific strain energy at any noncentroidal point and at the centroid of the element. Along a similar line, Shephard [19] has proposed an adaptive mesh refinement algorithm based on the strain energy density function and its variation of values within the finite element mesh. Again, the discretization error is measured

by the lack of uniformity in the strain energy density over the element volume.

Babuska and his associates have offered more mathematically rigorous, residual-based approaches to error analysis. Babuska and Rheinboldt [20] derived error bounds for the energy norm of the error based on the residual of the differential equation. Element error indicators were introduced as a means of determining which elements must be refined. Mesh optimality is based on achieving equal error indicators for all elements. Subsequent work continued to develop residual-based error estimators and associated adaptive mesh refinement schemes ([21],[22]).

Other residual-based methods include the work of Kelly, Gago, and Zienkiewicz ([23],[24]). In an attempt to reduce the computational burden involved in residual-based error estimates, the authors proposed the use of special hierarchical shape functions. The hierarchical shape functions essentially permit efficient p -based error estimates. However, a common fundamental problem of all residual-based methods has been the difficulty of correlating the residual, as measured by an integral norm, to the pointwise error in either the primary variable or its derivatives, such as stresses in elasticity.

More recently, Zienkiewicz and his colleagues ([25]-[27]) have derived a new stress-based error estimator and associated adaptivity algorithm. The method involves obtaining a global least squares fit of the discontinuous (C^0 continuous) finite element stress field with a piecewise continuous (C^1 continuous) stress field. The latter stress field is taken as an approximation to the true stress field, and the difference between the two tensor fields, as measured by the energy or L_2 norm, represents an estimate of the discretization error. The authors invoke the asymptotic convergence rate of displacement-based finite elements to correlate the norm of the error in the element stress field to the finite element size, thereby deriving an adaptive element sizing function for the new mesh.

Ainsworth et al [28] have shown that the Zienkiewicz and Zhu error estimator is effective and asymptotically exact provided that the exact stress boundary conditions are imposed on the higher order stress field. However, imposition of exact stress boundary conditions for general multi-dimensional problems, while straightforward, involves additional computations. Cauchy stress components must be transformed to boundary-based normal/tangential coordinate systems for all boundary finite elements. Moreover, the order of the system of equations generated by the least squares fit problem is, in general, larger than the original finite element system of equations. Thus, the algorithm is computational intensive, unless the coefficient matrix for the least square fit problem is diagonalized as recommended by Zienkiewicz and Zhu. The effect of this diagonalization or lumping of the coefficient matrix on the effectiveness, accuracy, and convergence properties of the algorithm was not been explored by Ainsworth et al.

4 *Posteriori* Error Estimators for h Refinement

To improve finite element based reliability predictions, the reliability engineer must be assured that the approximate finite element solution has met specified accuracy requirements. In general, there are three sources of error in finite element analysis:

- **truncation and roundoff error:** This error is inherent in any numerical method due to the limited precision in which numbers can be represented on digital computers. Normally, this error is insignificant in finite element analysis since computations are carried out in double precision and material properties, loads, and geometry are known with far less precision.
- **modeling error:** This error is due to improper modeling assumptions, such as assuming a fixed boundary condition when in fact the support is flexible. Practicing good engineering judgement and verifying results against assumptions can minimize modeling errors.
- **discretization error:** This is the error specific to the finite element method due to the representation of a continuous object and the field solution, such as the temperature distribution, in a piecewise fashion using a finite number of elements. Increasing the number of elements,

i.e. refining the mesh, decreases these errors, provided that proper elements are used. This error is often the major source of inaccuracies in finite element analysis and is often the most difficult to assess.

In this report we proposed and evaluated two new error estimators for measuring the finite element discretization error. The error estimators are based on formal mathematical techniques applied to the finite element analysis results. As such, they are termed *posteriori* error estimators. The new error estimators and corresponding h adaptive h -refinement algorithm are based on a generalized scalar energy density field, or in the case of elasticity, on the effective *von Mises*'s scalar stress field. The fundamental concept of these error estimators is similar to the error estimator proposed by Zienkiewicz and Zhu [25] in that the error estimator is based on the comparison of a discontinuous finite element quantity to an improved piecewise continuous (C^1 continuous) quantity. However, there are some important distinctions which we will elaborate later.

The following notation convention is used throughout this paper. All vector quantities are enclosed by curly braces $\{ \}$ and matrices by square brackets $[]$, all nodal variables are represented by an overbar, while finite element field variables, such as temperature distribution are indicated with an overhat. The subscript e denotes an element quantity.

Let $\{\hat{\sigma}\}$ denote the finite element stress tensor field which is, in general, discontinuous across interelement boundaries. Let $\{\sigma^*\}$ denote a C^0 continuous approximation to the true stress field. An estimate of the solution error is given by the difference of these two ~~scalar~~ fields. To obtain an expression for $\{\sigma^*\}$, two methods were examined: The Least Squares Fit (LSF) Method and a Statically Equivalent (SE) Method.

4.1 Least Squares Fit Method

To arrive at an expression for $\{\sigma^*\}$, Zienkiewicz and Zhu [25] imposed a weighted residual equality on the stress error estimate, $\{E\} \equiv \{\sigma^*\} - \{\hat{\sigma}\}$:

$$\int_V [N_\sigma]^T (\{\sigma^*\} - \{\hat{\sigma}\}) dV = \{0\} \quad (1)$$

where $[N_\sigma]$ is the matrix of interpolating polynomials associated with the $\{\sigma^*\}$ stress vector field. Following the standard finite element procedure used to formulate C^0 displacement fields, the C^0 stress field is expressed in terms of unknown nodal values $\{\bar{\sigma}^*\}$ using known nodal interpolating (shape) functions $[N_\sigma]$:

$$\{\sigma^*\} = [N_\sigma] \{\bar{\sigma}^*\} \quad (2)$$

Eq. (1) represents a system of $K \cdot N$ equations, where K equals the number of Cauchy stress components (i.e. 3 for 2-D, 6 for 3-D), and N is the number of nodes in the finite element model. The least squares interpretation of this approach can be seen as follows. Let the functional $\Pi(\{\sigma^*\})$ denote the squared L_2 norm of the error estimate:

$$\Pi(\{\sigma^*\}) = \int_V (\{\sigma^*\} - \{\hat{\sigma}\})^2 dV \quad (3)$$

By virtue of Eq. (2), Π is a function of $\{\bar{\sigma}^*\}$, the unknown nodal values of the projected stress field. By substituting for $\{\sigma^*\}$ from Eq. (2), setting the variation of Π to zero to minimize the error, and algebraically manipulating, one can derive Eq. (1).

From Eqns. (1) and (2), $\{\sigma^*\}$ can be found to be

$$\{\sigma^*\} = [N_\sigma][A]^{-1} \int_V [N_\sigma]^T \{\hat{\sigma}\} dV \quad (4)$$

where

$$[A] = \int_V [N_\sigma]^T [N_\sigma] dV \quad (5)$$

Zhu and Zienkiewicz have demonstrated the effectiveness of this stress based error estimator, and recently Ainsworth et Al. [28] have provided a deeper mathematical interpretation of the Zeinkiewicz and Zhu error estimator. However, there are certain disadvantages of this method, namely

- computationally expensive, unless a diagonal approximation for $[A]$ is employed.
- error norm estimates computed by the least squares method do not provide an upperbound on the true error norms.

The least squares method, if fully implemented, is computationally expensive because the order of the system of equations which must be factored and solved is $3N$ (2-D) or $6N$ (3-D), where N is the number of nodes. In contrast, the order of the original finite element system of equations is $2N$ and $3N$ for 2-D and 3-D problems, respectively. Thus, the computational effort required for computing an error estimate is actually greater than that required for computing the original finite element displacement solution.

The lack of upperboundedness of the error norm estimate to the true error norm is a direct consequence of the least squares fitting of the discontinuous finite element stress field. In general, in highly stressed regions the finite element solution underestimates the stress field. Thus, the least squares fit of the finite element solution does not improve the prediction of peak stresses, and the error estimator, as measured by an integral norm, is less than the true error norm. Moreover, the higher the discretization error, the greater this discrepancy between the estimated and true error norms. This condition will result in slower convergence (i.e. more mesh refinement iterations)

to an optimally refined mesh compared to that achieved by an error norm estimate which would upperbound the true error norm.

The computational efficiency of the error analysis can be significantly improved if the error analysis is based on a scalar function. In the following sections we present two new error estimators which are both based on a scalar function of the gradients in the field variables. The work is presented for elasticity problems in which the scalar function selected is the *von Mises* or effective stress. However, the method can be applied to any scalar function which, in the finite element solution, is discontinuous across interelement boundaries, and in which an improved scalar function is C^0 continuous. For example, in thermal analysis, a logical scalar function is the thermal energy density which is a function of the discontinuous heat fluxes.

In elasticity the effective or *von Mises* stress σ_v is a particularly attractive choice for four reasons:

1. It is based on the discontinuous finite element stress components.
2. A C^0 continuous σ_v field is generally an improved solution, although exceptions do exist.
3. All elasticity finite element codes must compute σ_v because of its failure-related significance.
4. The use of an failure-related stress function as a basis of error analysis permits accuracy requirements to be tightly integrated to the stress state.

The last point should be emphasized. The Distortional Energy Failure Theory is based entirely on *von Mises* stress function and is the most widely accepted failure theory for ductile materials. A error estimator based on the *von Mises* stress function enables the error estimator to have engineering significance, and permits the concept of adaptive accuracy to be introduced. Typically, an optimal mesh is defined as one which minimizes either the total error, as measured by an integral norm, or produces a minimum uniform local relative error distribution for a fixed number

of degrees of freedom. However, from a practical standpoint, an optimal mesh is one that offers a sufficiently accurate solution, as specified by the user, *in the critical regions of the design*, while allowing for less accurate solutions in less critical regions. With this definition of mesh optimality, a dramatic reduction in the number of degrees of freedom in the refined mesh can typically be achieved compared to a refined mesh in which the accuracy requirement is held constant throughout the model. Obviously, special care must be exercised for problems involving large deflections, buckling, eigenvalue extraction, etc. where local solution inaccuracies may have significant effects on global finite element quantities.

In 3-D, the *von Mises* stress is computed from the Cauchy stresses as follows:

$$\hat{\sigma}_v = \frac{1}{\sqrt{2}} \left[(\hat{\sigma}_x - \hat{\sigma}_y)^2 + (\hat{\sigma}_y - \hat{\sigma}_z)^2 + (\hat{\sigma}_z - \hat{\sigma}_x)^2 + 6 (\tau_{xy}^2 + \tau_{yz}^2 + \tau_{zx}^2) \right]^{1/2} \quad (6)$$

Since $\hat{\sigma}_v$ is a function of the discontinuous finite element Cauchy stress components, *von Mises* stress function will also exhibit interelement discontinuity. Thus, the discretization error E can be measured by $\hat{\sigma}_v - \sigma_v^*$, where σ_v^* is a piecewise continuous *von Mises* stress field. Zienkiewicz and Zhu's least squares method used to obtain the piecewise continuous stress tensor field $\{\sigma^*\}$ can also be applied to obtain an expression for σ_v^* :

$$\int_V [N_{\sigma_v}]^T (\sigma_v^* - \hat{\sigma}_v) dV = \{0\} \quad (7)$$

where

$$\sigma_v^* = [N_{\sigma_v}] \{\bar{\sigma}_v^*\} \quad (8)$$

It should be noted that Eq. (7) must be imposed independently on subdomains distinguished by distinct material properties. This is in accordance with the fact that the effective stress may in

reality be discontinuous across intermaterial boundaries. Thus, imposition of C^0 continuity across these boundaries would violate physical principles and weaken the accuracy of the error estimator.

Eq. (7) is solved to yield

$$\sigma_v^* = [N_{\sigma_v}][A]^{-1} \left(\int_v [N_{\sigma_v}]^T \hat{\sigma}_v dv \right) \quad (9)$$

where

$$[A] = \int_v [N_{\sigma_v}]^T [N_{\sigma_v}] dV \quad (10)$$

Note that since σ_v is a scalar function, the matrix $[A]$ to be inverted is of order N only. The Cholesky decomposition and forward and back substitution of a symmetric system of equations of order n involves $(n^2 + n)/2$ floating point operations (FLOPS) [29]. Thus, the ratio of FLOPS required for obtaining the stress vector field $\{\sigma^*\}$ to the FLOPS required for obtaining the scalar function σ_v^* is approximately K^2 , where K equals 3 for 2-D and 6 for 3-D problems. One may argue that this computational savings is offset to some extent by the FLOPS required to compute $\hat{\sigma}_v$ via Eq. (6). However, the engineering significance of σ_v dictates its computation irrespective of any *posteriori* error analysis.

If a diagonal approximation for $[A]$ is employed, then the FLOP ratio of the two methods is approximately K . However, the theoretical effect of a diagonal approximation to $[A]$ on the error estimator has yet to be explored.

Although the *von Mises* stress-based error estimator obtained by the least squares fit approach provides a measure of the discretization error and is computationally efficient, it possesses one notable disadvantage. It underestimates the peak *von Mises* stresses, and therefore the estimated

error norms underestimate the true error norms for most critical analysis regions of the finite element model. In the following section, we propose an alternative formulation for obtaining σ_v^* and an improved error estimator. The two approaches will be compared in the Examples section.

4.2 The Statically Equivalent C^0 Stress Field Method

This method was initially proposed by Loubignac et Al. [30] to improve the estimation of peak stresses. Modifications have been introduced by other researchers more recently ([31],[32]).

The static equilibrium equations are given by

$$[K]\{D\} - \{R\} = 0 \quad (11)$$

where

$$[K] = \int_V [B]^T [E] [B] dV \quad (12)$$

Thus

$$\left(\int_V [B]^T [E] [B] dV \right) \{D\} - \{R\} = 0 \quad (13)$$

We note that

$$\{\hat{\sigma}\} = [E][B]\{D\} \quad (14)$$

Therefore Eq. (13) also implies that

$$\int_V [B]^T \{\hat{\sigma}\} dV - \{R\} = 0 \quad (15)$$

Thus, the discontinuous finite element stress field $\{\hat{\sigma}\}$ satisfies static equilibrium by virtue of Eq. (15). If $\{\hat{\sigma}\}$ is replaced with a C^0 continuous stress field $^e\{\sigma^*\}$, Eq. (15) will no longer be satisfied

and a residual will exist due to the dissatisfaction of static equilibrium. Let

$$\{R_\sigma\} = \int_V [B]^T \{\sigma^*\} dV = \sum_{e=1}^m \int_{V_e} [B]_e^T \{\sigma^*\}_e dV \quad (16)$$

Comparing this with the applied load vector, the residual load vector $\{\Delta R\}$ is obtained. Thus, a Newton-Raphson iteration algorithm may be employed to update nodal displacements and finite element stresses until convergence, i.e. static equilibrium of the C^0 continuous $\{\sigma^*\}$ field, is achieved in some sense.

$$\{R_\sigma\}^i = \sum_{e=1}^m \int_{V_e} [B]_e^T \{\sigma^*\}_e^i dV \quad (17)$$

$$\{\Delta R\}^i = \{R\}^i - \{R_\sigma\}^i \quad (18)$$

$$[K] \{\Delta D\}^i = \{\Delta R\}^i \quad (19)$$

$$\{D\}^{i+1} = \{D\}^i + \{\Delta D\}^i \quad (20)$$

$$\{\hat{\sigma}\}_e^{i+1} = [E]_e [B]_e \{D\}_e^{i+1} \quad (21)$$

The C^0 continuous stress field $\{\sigma^*\}$ required in Eq. (17) may be obtained by simple nodal averaging of $\{\hat{\sigma}\}$ stresses (extrapolated from Gauss points) to obtain $\{\bar{\sigma}^*\}$ and then interpolating for $\{\sigma^*\}$ via Eq. (2). To adopt this method to our effective stress based error estimator, the $\{\sigma^*\}$ stress components can be used to directly compute σ_v^* via Eq. (6).

An appealing feature of this method is that the decomposed global matrix $[K]$ used in the original finite element analysis may be used for all Newton-Raphson iterations. Furthermore, this method has also been observed to overestimate peak stresses with the magnitude of peak stress overestimation being proportional to the coarseness of the mesh (i.e. higher discretization errors). Typically,

only a few iterations are needed before peak stresses are overestimated. Therefore, error estimations based on σ_v^* computed by this method tend to be conservative (i.e. overestimate the true error) for peak stress regions. This improves the convergence characteristics of the adaptivity algorithm in these critical regions.

5 Assessing the Discretization Error

Ideally, one would like to control the discretization error in a pointwise sense, since the ultimate objective of adaptive mesh refinement is to force the pointwise error in the finite element solution to be acceptable throughout the domain. However, in practice, norms must be used to measure the error element by element and modify the mesh accordingly.

The L_2 integral norm of a function f is given by

$$\|f(x, y, z)\|_2 \equiv \left[\int_V f^2(x, y, z) dV \right]^{\frac{1}{2}} \quad (22)$$

We note that the projected field σ_v^* is a C^0 continuous field which minimizes the L_2 norm of the difference between σ_v^* and $\hat{\sigma}_v$ over the entire domain. Therefore, a rather natural measure of the *estimated* discretization error for element e , $E(e\sigma_v)$ is:

$$\|E(e\sigma_v)\|_2 = \|\hat{\sigma}_v - e\sigma_v^*\|_2 = \left[\int_{eV} (\hat{\sigma}_v - e\sigma_v^*)^2 dV \right]^{1/2} \quad (23)$$

Note that we have slightly modified our notation, using the pre-superscript e , instead of the subscript e , to denote an element quantity.

If the L_2 norm is normalized by dividing by $V^{1/2}$, then it represents an integral root mean square

(RMS) value of the function. Thus, an RMS measure of the element's *von Mises* stress field is

$$RMS({}^e\sigma_v) = \left[\left(\int_{{}^eV} ({}^e\sigma_v^*)^2 dV \right) / {}^eV \right]^{\frac{1}{2}} \quad (24)$$

It is observed that the term in parenthesis in Eq. (24) is directly proportional to the element's distortional strain energy. Denoting the element's distortional strain energy as eDSE , we have the following relationships between the distortional strain energy, the RMS and L_2 norm of ${}^e\sigma_v$:

$${}^eDSE \propto {}^eV (RMS({}^e\sigma_v))^2 = \|{}^e\sigma_v\|_2^2 \quad (25)$$

Accordingly, the L_2 norm of the element's *von Mises* stress field is proportional to the square root of the element's distortional strain energy, and the RMS measure of the element *von Mises* stress field is proportional to the square root of the element's average distortional strain energy density.

Finally, a global RMS measure of the *von Mises* stress is given by

$$GRMS(\sigma_v) = \left[\left(\int_V (\sigma_v^*)^2 dV \right) / V \right]^{\frac{1}{2}} \quad (26)$$

$$= \left[\left(\sum_{e=1}^m \int_{{}^eV_e} ({}^e\sigma_v^*)^2 dV \right) / V \right]^{\frac{1}{2}} \quad (27)$$

$$= \left[\left(\sum_{e=1}^m {}^eV (RMS({}^e\sigma_v))^2 \right) / V \right]^{\frac{1}{2}} \quad (28)$$

where m equals the number of elements. As before, we identify the term in large parenthesis in Eq. (26) to be proportional to the total distortional strain energy ($TDSE$), yielding

$$TDSE \propto V (GRMS(\sigma_v))^2 = \|\sigma_v\|_2^2 \quad (29)$$

6 Adaptive Mesh Refinement

An adaptive mesh refinement scheme is simply an algorithm by which the results (i.e. stress or temperature distribution) of a finite element analysis can be used to remesh or refine the finite element model to improve the level of accuracy of the solution. Adaptive meshing techniques can be classified into three categories:

1. **r Refinement:** In this approach neither the connectivity of the finite element mesh or the order of the finite elements is changed. The new mesh contains the same number of nodes, elements, and nodal connectivity used to define each element. However, the nodal locations are moved to obtain a more optimal solution. The repositioning of the nodes may be based on a local error estimate or on the minimization of the total potential energy with respect to the nodal coordinates.
2. **h Refinement:** This is the most common type of mesh refinement scheme. It is often practiced in an intuitive manner by analysts without any formal error estimate computations. Ideally, error estimates are used to increase the number of elements and nodes locally and/or globally to obtain a more optimal mesh. The order of the element (linear, quadratic, etc.) remains unchanged.
3. **p Refinement:** This is a relatively new method where the number of elements and element shape remain the same. Instead, the order of the element is increased by the addition of midside and interior nodes. Higher order elements used higher order polynomials to represent the displacement or temperature field. The accuracy of the solution improves because of the additional nodal degrees of freedom in the new mesh.

Each method has its own advantages and disadvantages. In r refinement, the most optimal mesh can be obtained for a fixed number of degrees of freedom. However, there is no guarantee that the number of degrees of freedom in the mesh is sufficient to satisfy the accuracy requirements of the user. The h refinement method has the advantage of being applied locally and/or globally to force the solution error to within acceptable bounds. However, mesh restructuring is required which usually dictates the need for a robust automatic mesh generator. The p refinement method has shown promising results compared to the h refinement method. However, it is more difficult to

implement and requires access to the finite element source code which is not required by the other two techniques.

Recently, a powerful 2-D automatic mesh generation code has been developed at Sandia National Laboratory called FASTQ and currently installed on the Computer Aided Systems Engineering Vax computer at Rome Air Development Center. Due to the availability of this automatic 2-D mesh generator and due to the lack of source code access to commercial finite element codes, an h refinement adaptive meshing strategy was selected for this research.

The discretization error is assessed using element norms. For this we have employed the L_2 norm of the error defined in Eq. (23). Now, the strategy is to force this measure of the error to be sufficiently small relative to a reference value in order to achieve the desired pointwise accuracy in σ_v . The L_2 norm of the projected (i.e. assumed exact) von Mises stress field, $\|\sigma_v\|_2$ is taken as the reference value. Thus, the accuracy criteria is

$$\|\epsilon \hat{\sigma}_v - \epsilon \sigma_v^*\|_2 \leq \eta \|\sigma_v\|_2 \quad (30)$$

where η is small. However, difficulty lies in determining whether $\|\sigma_v\|_2$ norm should be computed locally or globally. If this norm is computed locally (i.e. over the element domain), then the adaptivity algorithm will be excessively demanding [25]. In effect, a uniform relative accuracy requirement will be imposed for all elements regardless of the magnitude of the element stress. The absolute estimated error may be very small for an element, but because the element is understressed, $\|\epsilon \sigma_v^*\|_2$ is also small, and the algorithm will demand refinement in spite of the small absolute error. On the other hand, if $\|\sigma_v^*\|_2$ in Eq. (30) is computed globally by summing up $\|\epsilon \sigma_v^*\|_2$ for all elements, then η loses its significance as the relative error norm ratio for each element.

In an attempt to resolve this dilemma, an adaptive accuracy criteria is introduced which can yield "optimal" meshes from a practical viewpoint. Let η denote a nominal target error fraction of the global L_2 norm of the C^0 stress field. Then, the following accuracy criteria is used:

$$\|E(\epsilon \sigma_v)\|_2 = \|\epsilon \hat{\sigma}_v - \epsilon \sigma_v^*\|_2 \leq \epsilon \eta (\|\sigma_v^*\|_2)_{\max} \quad (31)$$

where

$$\epsilon \eta = \eta \left(\frac{GRMS(\sigma_v^*)}{RMS(\epsilon \sigma_v^*)} \right)^\alpha \quad (32)$$

$$(\|\sigma_v^*\|_2)_{\max} = \max_{e=1}^m \|\epsilon \sigma_v^*\|_2 \quad (33)$$

where α is a constant ≥ 0 which controls the degree of accuracy adaptivity. If $\alpha = 0$, then a nonadaptive accuracy requirement is imposed. It is observed that the effect of Eq. (32) is to demand the greatest accuracy for the element with the highest RMS measure of σ_v^* and to relax the accuracy requirement for all elements where $RMS(\epsilon \sigma_v^*) < GRMS(\sigma_v^*)$. Practical bounds must be placed on $\epsilon \eta$ computed by Eq. (32) to prevent mesh transition problems. For example, $\epsilon \eta$ may be permitted to vary only from 0.01 to 0.20 with a nominal value of 0.05.

Note we have selected the maximum element L_2 norm of σ_v^* as the reference value in Eq. (31) to which the element norm of the error, $\|E(\epsilon \sigma_v)\|_2$, are compared. This choice represents a compromise between a purely element-based reference value and a purely global-based reference value. Numerous numerical experiments based on local, global, and local/global reference values indicated this choice avoided the problem of excessive mesh refinement in understressed regions, while retaining the ability of the mesh to converge quickly to sufficiently accurate solutions in highly stressed regions. In a similar manner we have found that an accuracy adaptivity constant of $\alpha \approx 2.0$ yields good results in terms of adapting the accuracy of the solution to the rate of change of the stress field. However, more work is clearly needed in this area.

While the error norm ratio $\frac{\|E(\epsilon\sigma_v)\|_2}{(\|\sigma_v^*\|_2)_{max}}$ will not bound the pointwise error $E(\sigma_v)$, the pointwise error will generally converge faster than the error norm ratio as the mesh is refined. Thus, the strategy is to use the error norm ratios as a basis for adaptivity, while monitoring the relative estimated pointwise errors in σ_v at the peak stress points in the mesh.

An h -Based Adaptive Remeshing Scheme:

Let $\bar{\epsilon}$ denote the right hand side of Eq. (31). Define the error ratio ξ as

$$\xi_e = \frac{E(\epsilon\sigma_v)}{\bar{\epsilon}} = \frac{\|\epsilon\sigma_v^* - \epsilon\hat{\sigma}_v\|}{\bar{\epsilon}} \quad (34)$$

If $\xi_e > 1$, then the element size h_e is decreased. If $\xi_e < 1$, then we increase the element size h_e . If $\xi = 1$ for all elements, then the desired adaptive accuracy requirement has been achieved.

To obtain appropriate values by which to increase or decrease element size h , the asymptotic convergence rate criteria is invoked at the element level. If the order of the assumed polynomial used to approximate the field variable is p , then one can expect a convergence rate of the field variable of order h^{p+1} since higher order terms have been omitted from the assumed field. Similarly, the convergence rate for derivatives of the field variable, such as strains or temperature gradients, will have a convergence rate of order h^{p+1-l} , where l is the order of differentiation required to obtain the derivative variables. One can expect, then, that the convergence rate of $E(\epsilon\sigma_v)$ will be of order h^{p+1-l} . Since the L_2 norm of $E(\epsilon\sigma_v)$ is given by the square root of the integral of $E(\epsilon\sigma_v)$, it is argued that the convergence rate for $\|E(\epsilon\sigma_v)\|_2$ is

$$CR(\|E(\epsilon\sigma_v)\|_2) \propto [\int (h^{p+1-l})^2 dV]^{1/2} = (h^{2(p+1-l)+1})^{1/2} \quad (35)$$

For unear elements and 2D and 3D elasticity, $p = 1$ and $l = 1$, yielding

$$CR(\|E(\epsilon\sigma_v)\|_2) \propto h^{3/2} \quad (36)$$

Denoting the current mesh with the pre-subscript i , the requirement that $\epsilon_{i+1}\xi = 1$ with Eq. (36) yields

$$\frac{\epsilon_i \xi}{\epsilon_{i+1} \xi} = \epsilon_i \xi = \frac{\epsilon_i h^{(2p+1)/2}}{\epsilon_{i+1} h^{(2p+1)/2}} \quad (37)$$

Thus,

$$\epsilon_{i+1} h = \epsilon_i \xi^{-2/(2p+1)} \epsilon_i h \quad (38)$$

For linear elements, one has

$$\epsilon_{i+1} h = \frac{\epsilon_i h}{\epsilon_i^{2/3}} \quad (39)$$

7 Automatic Adaptive h -refinement Mesh Generation

Recently, a new automatic mesh generation technique has been developed at Sandia National Laboratories which is ideally suited for adaptive h -refinement. The technique, called "paving" in 2-D and "plastering" in 3-D, uses a boundary description of the geometry and user-specified element sizing functions along the boundary segments to generate all quadrilateral meshes. Essentially, rows of elements are laid down, element by element, in a counter clockwise direction around the exterior perimeter preceding inward and in clockwise directions around interior perimeters (i.e. holes) preceding outward. Sophisticated algorithms are employed to transition between different element sizes and to correct the mesh where rows of elements overlap. The meshing algorithms are implemented in a large FORTRAN code called FASTQ (FAST Quadrilaterals). The reader is referred to Reference [33] for more details of this powerful mesh generation technique for quadrilateral meshes.

Because elements are inserted into the mesh one by one, the algorithm is well suited for adaptive meshing. Details of this algorithm will be forthcoming.

8 Implementation

The two error analysis approaches of the previous section were implemented into two separate in-house finite element codes, called **ERRFEM1** and **ERRFEM2**. The **FASTQ** 2-D automatic mesh generation code, discussed above, was modified to implement adaptive meshing according to Eq. (39). Subroutines were also added to **FASTQ** for automatic generation of complete input data files for **ERRFEM1** and **ERRFEM2** based on user-supplied data through interactive session.

Note that the error estimation algorithms have been implemented into in-house finite element codes, instead of implemented as separate stand-alone codes which interface to commercial finite element codes as originally planned (Fig. 2). This was necessary due to the technical difficulties of interfacing to a commercial code, the research nature of this work, and the time constraints on the project.

The sequence of activities with the error analysis and automatic adaptive mesh generation is illustrated in Fig. 3. First, the user defines an initial boundary representation of the 2-D geometry with associated meshing data either interactively in **FASTQ** or through the preparation of a **FASTQ** input data file. Meshing data consists of simply the type of element (four-, eight-, or nine-noded), the number of elements along each given boundary segment, the uniformity of element spacing along each boundary segment, and the type of mesh generation technique (i.e. paving technique). Next, **FASTQ** uses this information and the paving algorithm to automatically generate an initial

finite element mesh. The user then interactively requests that an input file be generated for the finite element/error analysis codes. The system then prompts the user for additional finite element modeling data, such as material properties, and loading and boundary conditions. After specification, the input file for **ERRFEMx** is written, as well as a **FASTQ** binary mesh database file and a **FASTQ** input file.

The finite element/error analysis code is then executed. Execution causes the input file to be read, a finite element analysis to be performed, error ratios to be calculated for each element, and the error ratio results to be appended to the end of the binary mesh database file. **FASTQ** is invoked, and the user selects the remeshing option under the mesh processing module. The system then prompts the user for the name of the **FASTQ** input data file and the binary mesh database file. The paving mesh generation algorithm is executed where new elements are sized based on the error ratios contained in the database file and old element sizes according to Eq. (39). The geometry definition information contained in the **FASTQ** input data file is used to ensure that remeshing is based on the originally defined geometry, and not the geometry defined by the previous mesh.

Note that this process could have been streamlined considerably by merging the **FASTQ** code with the **ERRFEM** codes. However, the resulting loss of modularity would have hindered future applications involving stand alone commercial finite element codes, stand alone error analysis codes, and stand alone automatic adaptive mesh generation. For this reason it was decided to keep adaptive mesh generation and finite element and error analysis separate. Furthermore, work is currently underway on implementing the error analysis algorithms in a stand alone code which will interface to **FASTQ** for mesh generation and **NISA2** for finite element analysis.

9 Examples

A comparison of the proposed error estimators and h -based adaptivity algorithms is illustrated with the following 2-D examples involving four-noded bilinear quadrilateral elements. The first example presented is the 2-D plane stress problem of a thin plate with hole under tension (Fig. 4(a)). The second example studied is the endloaded cantilever beam shown in Fig. 4(b).

The initial coarse-mesh finite element model representing one fourth of the plate is shown in Fig. 5(a), with the corresponding nodal averaged $\hat{\sigma}_v$ stress distribution shown in Fig. 5(b), where the stress distribution has been nondimensionalized by the applied traction σ_0 . The relative pointwise error (RPWE) is defined as

$$RPWE \stackrel{def}{=} \frac{((\sigma_v)_{\text{exact}} - \hat{\sigma}_v)}{(\sigma_v)_{\text{exact}}} \quad (40)$$

and has a value of 23.5% at the critical stress point (top of hole) for this mesh. The "exact" von Mises stress value (4.35) was obtained using a fine mesh consisting of 1400 elements, 1491 nodes, and 2940 active degrees of freedom. A series of convergence studies established that this mesh was sufficiently accurate to represent the "exact" stress distribution.

In Figures 5(c)-4 results are presented for the least squares fit (LSF) method for obtaining σ_v^* . A nominal target accuracy ratio of $\eta = 0.05$ was used with upper and lower limits of 0.01 and 0.20 placed on the adaptive accuracy ratio ${}^e\eta$. In Figs. 5(c) we show the distribution of adaptive accuracy ratio given by Eq. 32. The corresponding error ratio distribution given by Eq. (34) are plotted in Fig. 5(d). The maximum error ratio is 4.86 corresponding to ${}^e\eta = 0.014$ and occurs at the critical stress element.

Employing the new element size adaptivity function given by Eq. (29) and the "paving" mesh generation technique, the first mesh refinement is shown in Fig. 6(a). To avoid possible mesh transitioning problems, elements with error ratios less than one were not allowed to increase in size. Again, the *von Mises* stress distribution is shown in Fig. 6(b), and the relative pointwise error at the critical stress point is now 0.10. Adaptive accuracy distribution and corresponding error ratio distribution are shown in Figs. 6(c) and 6(d). Note that the maximum error ratio has been reduced from 4.86 to 3.30, while the relative pointwise error in σ_v at the top of hole has decreased from 23.5% to 10.0%. As expected, the pointwise error converges faster than the norm of the error.

Figs. 7(a)-7(d) show the third and final remesh results for the LSF method. The relative pointwise error at the critical stress point and the maximum error ratio are now 0.045 and 1.40, respectively.

Analogous results for the statically equivalent (SE) *von Mises* stress based method are shown in Figs. 8-10. Again, a target accuracy of $\eta = 0.05$ was sought. Six Newton-Raphson iterations were performed for each analysis to obtain the statically equivalent piecewise continuous *von Mises* stress field. The results indicate improved convergence characteristics of this method, as evidenced by the higher maximum error ratios for the same coarse mesh and faster convergence of the pointwise percentage error at the critical stress point.

The effect of adaptive accuracy factor α in Eq. (32) was explored by setting $\alpha = 0$ and repeating the analysis using the statically equivalent stress based method. The result was poor remeshes due to distorted elements. The error ratios were found to be higher at the distorted elements than at the critical stress region. Thus, refinement occurred at understressed but distorted element regions initially. After several remeshes, refinement did proceed to the critical stress region, albeit slowly,

but the target accuracy was not reached before the error ratio converged to one. The implication of this observation will be discussed in the following section.

Results for the plate with hole problem are summarized in Fig. 11 which illustrates the convergence characteristics of both the pointwise error and the error norm ratio of the two methods. From Fig. 11(a) one can see the superior convergence characteristics of the SE method compared to the LSF method in both the error norm ratio (ENR) for the critical element and the relative pointwise error (RPWE) at the critical stress point. In Fig. 11(b) the effect of adaptive accuracy on convergence is clearly demonstrated. The nonadaptive approach exhibited relatively slow convergence characteristics. The increasing slope of both the ENR and RPWE curves for the adaptive accuracy approach indicates that accuracy is increasing at a faster rate than the rate degrees of freedom are increasing, a phenomenon attributable to the localized mesh refinement in the vicinity of the critical stress point.

In the beam example only the SE method was used to obtain the improved σ_v^* stress field. The initial uniform finite element mesh for the beam problem is shown in Fig. 12(a). For convenience Poisson ratio was set to zero. The *von Mises* stress distribution, nondimensionalized by the applied traction σ_0 is shown in Fig. 12(b). At the critical stress point at the left edge on the top or bottom of the beam, the relative pointwise error is 11.3% for this coarse mesh. Figs. 12(c) and 12(d) show the adaptive accuracy and error ratio distributions obtained. Figures 13 and 14 present the results for the first and third (final) remeshes. These results are summarized in Fig. 15. Note the excellent convergence of the pointwise error to the target accuracy in the *first* remesh.

10 Discussion

In the plate example the error norm ratios for the critical stress element were found to be less than the relative pointwise error at the critical stress point for a given mesh. The opposite was true for the beam problem. This result is to be expected since the error norm does not bound the pointwise error. Since the norm of the error reflects an "average" error over the element volume, one can expect in areas of high stress concentration, such as at the top of the hole in the plate, the relative pointwise error to exceed the error norm ratio. Therefore, convergence of the error ratio to one is not a sufficient condition for obtaining the target accuracy in a pointwise sense. In practice we have observed that imposing an adaptive accuracy requirement according to Eq. 32 has proven to be effective in driving the critical pointwise error to be within the target accuracy as the error ratio approaches unity. However, more theoretical and numerical work is needed to provide an improved correlation in some sense of the relative pointwise error to an error norm ratio.

Two interesting observations are noted for the beam problem. First, by comparing Figs. 13(a), 13(c), and 13(d), it is clear that the SE method predicts a relatively high error ratio for distorted elements. Secondly, the final remesh shows relatively small element sizes on the beam's neutral axis where the y boundary condition constraint is imposed. Although stresses are not high at this point, a theoretical discontinuity exists in the σ_y stress across the interelement boundary at this point due to the reaction force applied by the boundary condition constraint. The result is a theoretical discontinuity in the *von Mises* stress field at this point. Therefore, the smoothed continuous *von Mises* stress field at this point does not represent an improved solution. Future work must account for such situations if stress based error estimators are to be reliably used for automatic adaptive mesh refinement.

11 Conclusions

Two new *von Mises* or effective stress based error estimators were introduced and compared. An effective stress based error estimator offers some significant advantages over residual and other stress-based error estimators. First, because it is a scalar function, the error estimator is computationally efficient to compute. Second, it provides a natural means by which the concept of adaptive accuracy can be introduced, thereby permitting optimal meshes from an engineering or failure-oriented perspective. Third, because the error estimator is based on a scalar function, the method is easily extendable to 3-D problems. Fourth, the method can be easily generalized for other problem domains by using similar scalar functions of engineering significance. For example, in thermal analysis, the thermal energy density function is a natural scalar function to use as a basis for error estimation.

Both methods demonstrated convergence of the critical pointwise error for the test case problems studied. The statically equivalent error estimator method was seen to offer superior convergence characteristics than the least squares fit method. However, for a given mesh the SE method is computationally more expensive than the least squares fit method, especially if a local least squares fit method is employed by simply averaging element stresses extrapolated to nodes. Thus, the superior convergence characteristics of the SE method is at the expense of increase computational costs for a given mesh. However, for most applications, it is preferable to minimize the number of mesh refinements needed because of the real time involved and computer resources needed to manage multiple finite element models. In view of this concern, the SE method is more attractive than the LSF method. A judicious combination of the two methods may afford an optimal path to a sufficiently accurate solution.

12 Future Work

The finite element and error analysis codes **ERRFEM1** and **ERRFEM2** are useful research tools for quickly implementing and testing error estimation algorithms. However, to be of greater utility to the Air Force, the error estimation algorithms must also have the capability to accept finite element results obtained from commercial finite element codes such as **NISA2**. In addition, the two error estimation algorithms should be integrated into a single code. Thus, the first task on the agenda is to bring **ERRFEM1** and **ERRFEM2** error analysis codes together into a single finite element/error analysis computer program which will interface to **FASTQ** data files for mesh data and to **NISA2** output files for analysis data, if desired. This task is already underway and is 70% completed in the form of a C computer code. The code is scheduled to be completed by May 1, 1991.

Future work will also include extension of the error estimation algorithms to thermal analysis problems. Extension to thermal analysis is critical since thermally induced stresses and strains often result in microelectronic device reliability problems. Thermal analysis error estimation can be achieved by replacing the *von Mises* stress function appearing in the algorithms with a scalar function, such as thermal energy density, of the heat flux vector. However, since heat flux is typically not computed at element gauss points by commercial finite element codes, the error analysis code will be modified to compute the heat flux distribution from finite element nodal temperature results. Thermal finite elements also must be added to the code so that the code will have a stand-alone thermal finite element analysis capability. This implementation is quite straightforward.

Error estimation based on comparing an "improved" piecewise continuous function (i.e. *von Mises* stress) with the finite element discontinuous function are based on a continuity assumption. It is

theoretical solution may, in fact, be discontinuous. One example is the *von Mises* stress function across intermaterial boundaries. Under these conditions, the in-plane Cauchy stress components parallel to the intermaterial boundary may be discontinuous, which results in a theoretically discontinuous σ_v stress function. Accordingly, one of the major focus of future work will be to modify the error estimation algorithm to handle intermaterial boundaries and other sources of stress discontinuities (such as point loads perpendicular to interelement boundaries). Conceptually, the proposed error estimation algorithms can be easily modified to account for these circumstances. However, implementation presents a more serious challenge. Since intermaterial boundaries are quite common in microelectronic devices, this work is also of critical importance.

Thirdly, the work must be extended to nonlinear problems. Again, this appears to be a straightforward task, because the error estimation algorithm can be applied independent of time and loading history. However, for path dependent nonlinear problems, the effect of small inaccuracies in the solution at previous time steps may have a significant impact for the current time step. This raises the question of conditional stability and its relationship to accuracy which must be explored.

Finally, future work will include extension of the error estimation and adaptive remeshing scheme to 3-D problems. A 3-D version of the FASTQ automatic mesh generation code is scheduled to be released in September 1991. Only minor modifications are needed to extend the error estimation algorithm to 3-D. However, it is expected that a significant amount of numerical studies will be needed to study, detect, and correct automatic adaptive mesh generation problems in 3-D.

References

- [1] W. J. Bocchi, "Finite element modeling and thermal simulations of transistor integrated circuits," In-House Report No. RADC-TR-89-176, Rome Air Development Center, NY, Oct. 1989.
- [2] W. J. Bocchi, J. A. Collins, and D. J. Holzhauer, "Thermal Stress Analysis of Integrated Circuits Using Finite Element Methods," In-House Report No. RADC-TR-84-100, Rome Air Development Center, NY, April 1984.
- [3] B. Romer and H. Pape, "Stress effects of package parameters on 4 Mega DRAM with fractional factorial designed finite element analysis," *Proc. of 39th. Electronic Components Conference*, May 1989, pp 832-839.
- [4] J. C. Glaser and M. P. Juare, "Thermal and structural analysis of a PLCC device for surface mount processes," *Journal of Electronic Packaging*, vol 3, no. 3, Sep 1989, pp 172-178.
- [5] K. R. Thompson, "Solder joint strain reduction of leadless IC packages through finite element analysis tailored PCB construction," *Sixth IEEE/CHMT Electronic Manufacturing Technology Symp.*, 1989, pp 105-111.
- [6] J. H. Lau, "Thermal stress analysis of SMT PQFP packages and interconnections," *Journal of Electronic Packaging*, vol 3, no. 1, Mar 1989, pp 2-8.
- [7] R. Nalbandian, K. La Rosa and K. Burton, "Automatic reliability assessment of surface mount components by finite element analysis," *9th International Electronics Packaging Conference*, vol 1, 1989, pp 555-571.
- [8] H. Hardisty and J. Abboud, "Thermal Analysis of Dual-in-line package using finite element method," *IEE Solid-state and Electron Devices Proc.*, vol 134, no. 1, Feb. 1987, pp 23-31.
- [9] J. A. Cooke and S. W. Lee, "Finite element thermal analysis of 144 pin plastic flat packs," *Fifth IEEE Semiconductor Thermal and Temperature Measurement Symp. Proc.*, Feb. 1989.
- [10] B. R. Simon, Y. Yuan, J. R. Umaretiya, J. L. Prince and Z. L. Staszak, "Parametric study of a VLSI plastic package using locally refined finite element models," *Fifth IEEE Semiconductor Thermal and Temperature Measurement Symp. Proc.*, Feb. 1989.
- [11] J. H. Lau and C. G. Harkins, "Thermal stress analysis of SOIC packages and interconnections," *IEEE Transactions on Components, Hybrids and Manufacturing Technology*, vol 11, no. 4, 1988, pp 380-389.
- [12] S. Kawai, "Structural design of plastic IC packages," *JSME International Journal, Series 1-Solid Mechanics, Strength of Materials*, vol 32, no. 3, July 1989, pp 320-330.
- [13] P. A. Engel and C. K. Lim, "Stress analysis in electronic packaging," *Finite Element Analysis and Design*, vol 4, no. 1, June 1988, pp 9-18.

- [14] J. H. Lau, D. W. Rice and P. A. Avery, "Elastoplastic analysis of surface mount solder joints," *IEEE Transactions on Components, Hybrids and Manufacturing Technology*, vol CHMT-10, no. 3, Sep 1986, pp 346-357.
- [15] M.S.Shephard, "Finite Element Grid Optimization- A Review ," *Finite Element Grid Optimization*, M.S.Shephard and R.Gallagher. eds., ASME Special Publication PVP-38, N.Y., 1979.
- [16] I.Babuska, O.C.Zienkiewicz, J.Gago, and E.R.A.Oliveira, *Accuracy Estimates and Adaptive Refinements in Finite Element Computations*, John Wiley & Sons, New York, 1986.
- [17] D.J.Turcke and G.M.McNeice, "Guidelines for Selecting Finite Element Grids Based on an Optimization Study," *Computers and Structures*, Vol. 4, 1974, pp. 499-519.
- [18] R.J.Melosh and P.V.Marcal, "An Energy Basis for Mesh Refinement of Structural Continua," *International Journal for Numerical Methods in Engineering*, Vol. 11, pp. 1083-1091, 1977.
- [19] M.S.Shephard, "Finite Element Grid Optimization with Interactive Computer Graphics," Ph.D. Thesis, Dept. of Structural Engineering, Cornell University, Jan. 1979.
- [20] I.Babuska, and W.C.Rheinboldt, "A-Posteriori Error Estimations for The Finite Element Method," *International Journal for Numerical Methods in Engineering*, Vol. 12, pp.1597-1615, 1978.
- [21] W.C.Rheinboldt, "Error Estimates for Nonlinear Finite Element Computations", *Computers & Structures*, Vol. 20, No. 1-3, pp. 91-98, 1985.
- [22] I.Babuska, and A.Miller, "The Post-processing Approach in the Finite Element Method - Part 3: A Posteriori Error Estimates and Adaptive Mesh Selection", *International Journal for Numerical Methods in Engineering*, Vol. 20, pp.2311-2324, 1984.
- [23] D.W.Kelly, J.P.Gago, O.C.Zienkiewicz, and I.Babuska, "A Posteriori Error Analysis and Adaptive Processes in the Finite Element Method: Part I, Error Analysis", *International Journal for Numerical Methods in Engineering*, Vol. 19, pp. 1593-1619, 1983.
- [24] J.P.Gago, D.W.Kelly, O.C.Zienkiewicz, and I.Babuska, "A Posteriori Error Analysis and Adaptive Processes in the Finite Element Method: Part II, Adaptive Mesh Refinement", *International Journal for Numerical Methods in Engineering*, Vol. 19, pp. 1621-1656, 1983.
- [25] O.C.Zienkiewicz, and Z.Zhu, "A Simple Error Estimator and Adaptive Procedure for Practical Engineering Analysis", *International Journal for Numerical Methods in Engineering*, Vol. 24, pp.337-357, 1987.
- [26] J.Z.Zhu, and O.C.Zienkiewicz, "Adaptive techniques in the finite element method," *Comm. Appl. Num. Methods*, Vol. 4, 197-204, 1988.
- [27] O.C.Zienkiewicz, and R.L.Taylor (1989), *The Finite Element Method* 4th Ed., McGraw-Hill, New York, pp. 398-435.

- [28] M.Ainsworth, J.Z.Xhu, A.W.Craig, and O.C.Zienkiewicz, "Analysis of the Zienkiewicz-Zhu A-Posteriori Error Estimator in th Finite Element Method," *International Journal for Numerical Methods in Engineering*, Vol. 28, 2161-2174, 1989.
- [29] G.H.Golub, and C.F.Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, 1983.
- [30] G.Loubignac, G.Cantin, and G.Touzot, "Continuous Stress Fields in Finite Element Analysis," *AIAA JOURNAL*, Vol. 15, No. 11 , pp. 1645-1647, 1977.
- [31] R.D.Cook, and X.Huang, "Continuous Stress Fields by the Finite Element -Difference Method," *International Journal for Numerical Methods in Engineering*, Vol. 22, pp.229-240, 1986.
- [32] R.D.Cook, D.S.Malkus, and M.E.Plesha, *Concepts and Applications of Finite Element Analysis*, John Wiley & Sons, New York, 1989.
- [33] T.D.Blackner, and M.B.Stephenson, "Paving: A New Approach to Automated Quadrilateral Mesh Generation," *Sandia Report*, SAND90-0249 .UC-705, 1990.

Figure Captions

Fig. 1: Reliability Assessment by the Finite Element Method

Fig. 2: Proposed Research Project

Fig. 3: Error Analysis Process Using FASTQ and ERRFEM Codes

Fig. 4(a): Plate with a Hole

Fig. 4(b): Cantilever Beam Under Bending

Fig. 5(a): Initial Coarse Mesh (16 Elements)

Fig. 5(b): Normalized *von Mises* Stress Distribution

Fig. 5(c): Adaptive Accuracy Ratio Distribution, LSF Method

Fig. 5(d): Error Ratio Distribution, LSF Method

Fig. 6(a): First Remesh, LSF Method (29 Elements)

Fig. 6(b): Normalized *von Mises* Stress Distribution

Fig. 6(c): Adaptive Accuracy Ratio Distribution, LSF Method

Fig. 6(d): Error Ratio Distribution, LSF Method

Fig. 7(a): Third (Final) Remesh, LSF Method (72 Elements)

Fig. 7(b): Normalized *von Mises* Stress Distribution

Fig. 7(c): Adaptive Accuracy Ratio Distribution, LSF Method

Fig. 7(d): Error Ratio Distribution, LSF Method

Fig. 8(a): Adaptive Accuracy Ratio Distribution, SE Method

Fig. 8(b): Error Ratio Distribution, SE Method

Fig. 9(a): First Remesh, SE Method (45 Elements)

Fig. 9(b): Normalized *von Mises* Stress Distribution

Fig. 9(c): Adaptive Accuracy Ratio Distribution, SE Method

Fig. 9(d): Error Ratio Distribution, SE Method

Fig. 10(a): Second Remesh, SE Method (81 Elements)

Fig. 10(b): Normalized *von Mises* Stress Distribution

Fig. 10(c): Adaptive Accuracy Ratio Distribution, SE Method

Fig. 10(d): Error Ratio Distribution, SE Method

Fig. 11(a): Convergence Comparison of LSF and SE Methods

Fig. 11(b): Convergence Comparison of Nonadaptive vs. Adaptive Accuracy

Fig. 12(a): Initial Coarse Mesh, SE Method (40 Elements)

Fig. 12(b): Normalized *von Mises* Stress Distribution

Fig. 12(c): Adaptive Accuracy Ratio Distribution

Fig. 12(d): Error Ratio Distribution

Fig. 13(a): First Remesh, SE Method (74 Elements)

Fig. 13(b): Normalized *von Mises* Stress Distribution

Fig. 13(c): Adaptive Accuracy Ratio Distribution

Fig. 13(d): Error Ratio Distribution

Fig. 14(a): Final (Third) Remesh, SE Method (168 Elements)

Fig. 14(b): Normalized *von Mises* Stress Distribution

Fig. 14(c): Adaptive Accuracy Ratio Distribution

Fig. 14(d): Error Ratio Distribution

Fig. 15: Convergence of the Error for the Beam Example

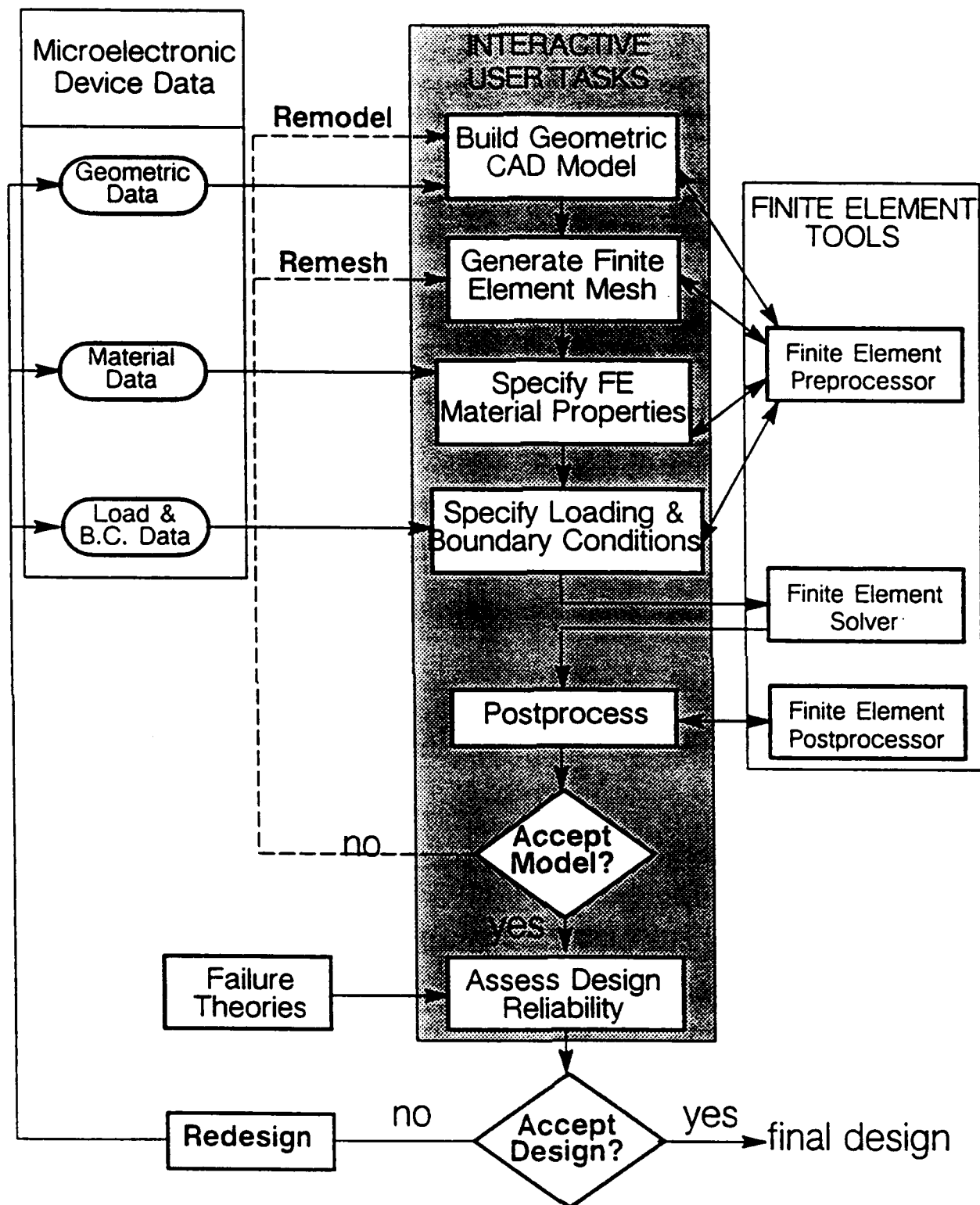


Figure 1: Reliability Assessment by the Finite Element Method

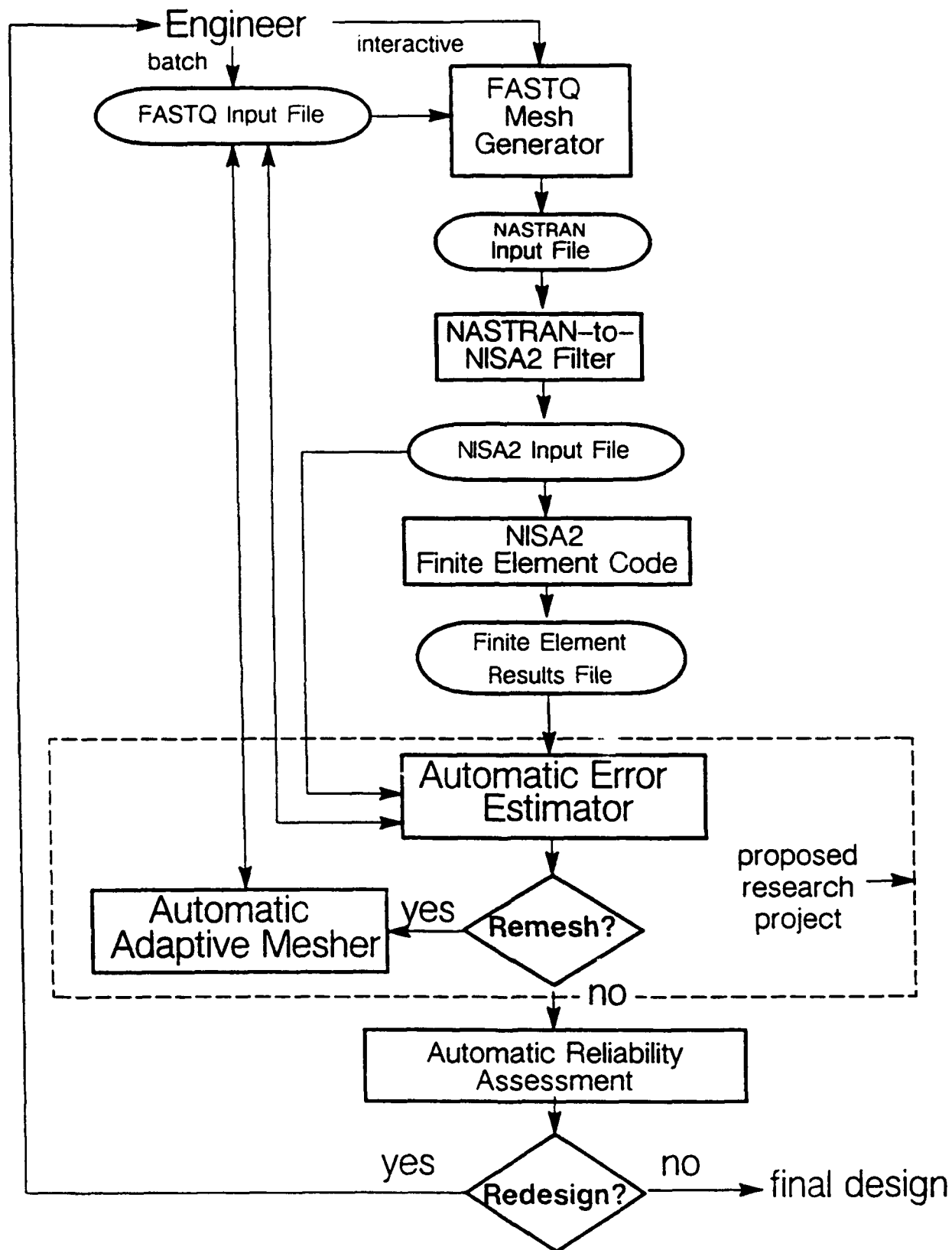
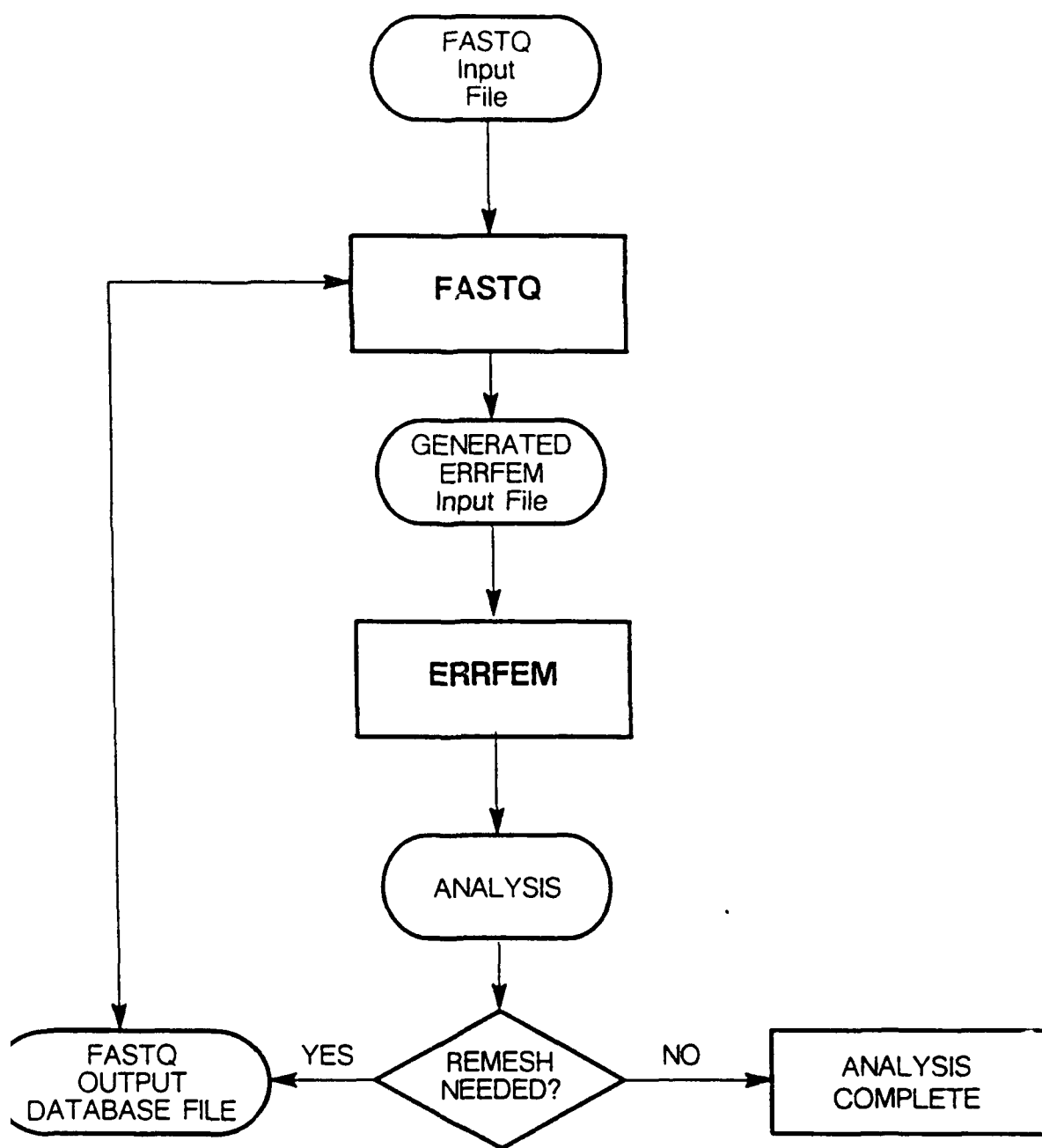


Figure 2: Proposed Research Project



ig. 3: ERROR ANALYSIS PROCESS USING FASTQ AND ERRFEM CODES

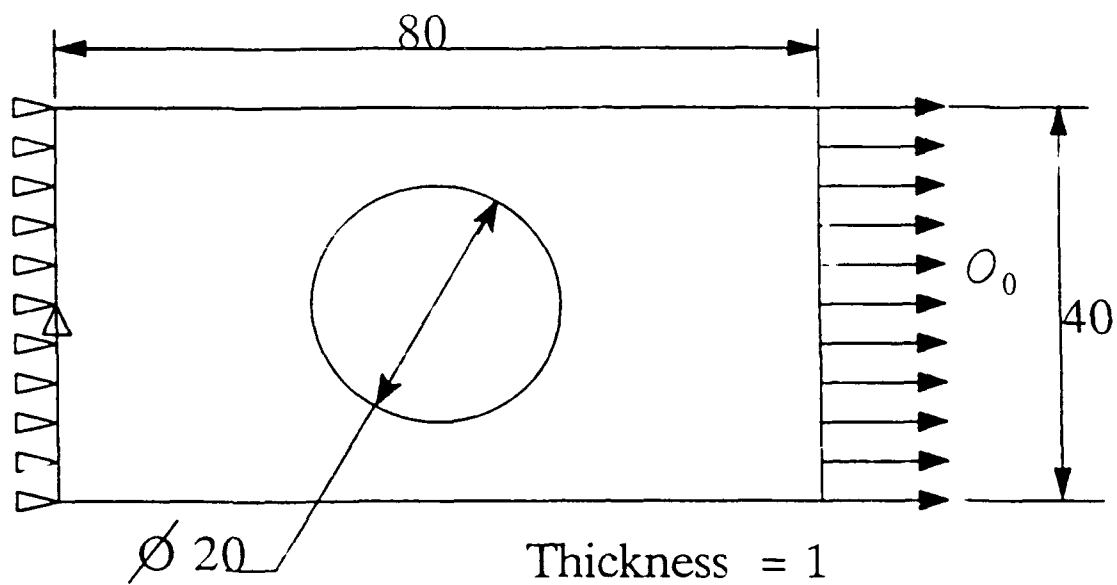


Fig. 4(a) Plate with a Hole.

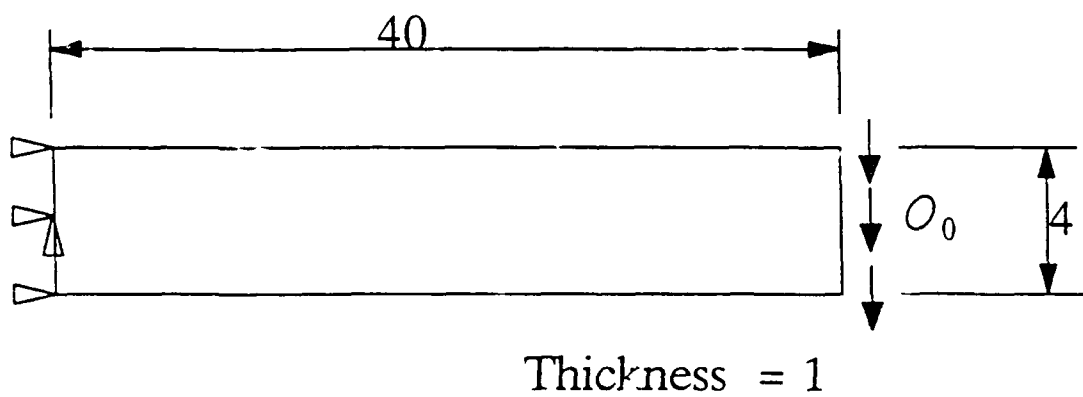


Fig. 4(b) Cantilever Beam Under Bending.

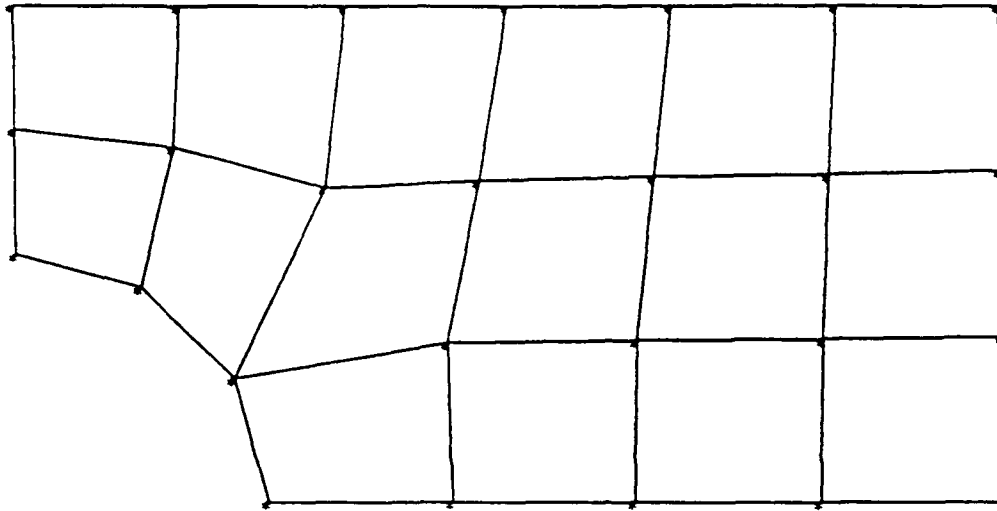
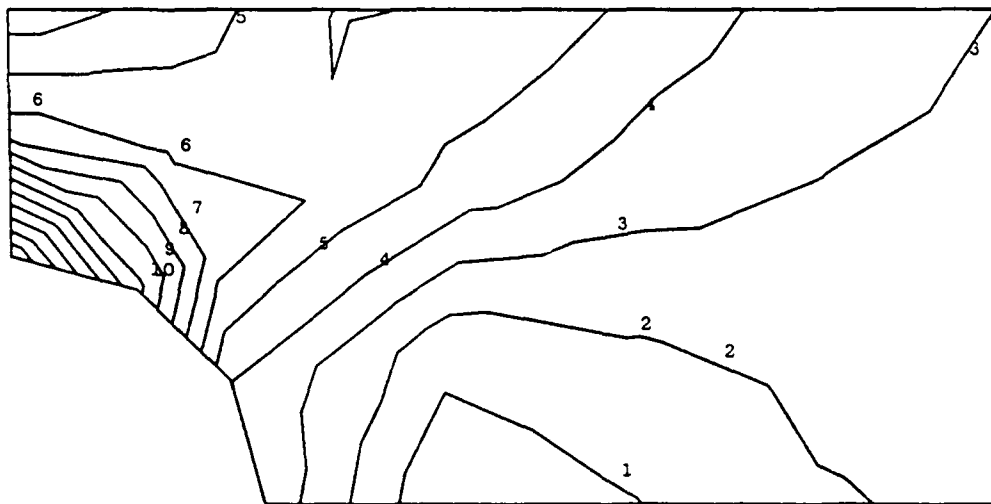


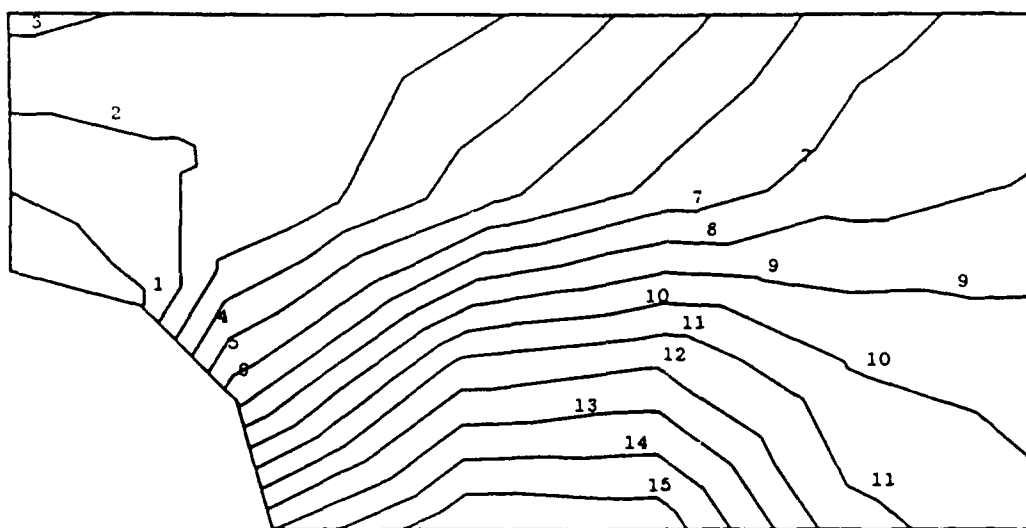
Fig. 5(a) Initial Coarse Mesh (16 Elements).



Min: 0.49
Max: 3.33

3.15	15
2.97	14
2.80	13
2.62	12
2.44	11
2.26	10
2.09	9
1.91	8
1.73	7
1.55	6
1.38	5
1.20	4
1.02	3
0.844	2
0.666	1

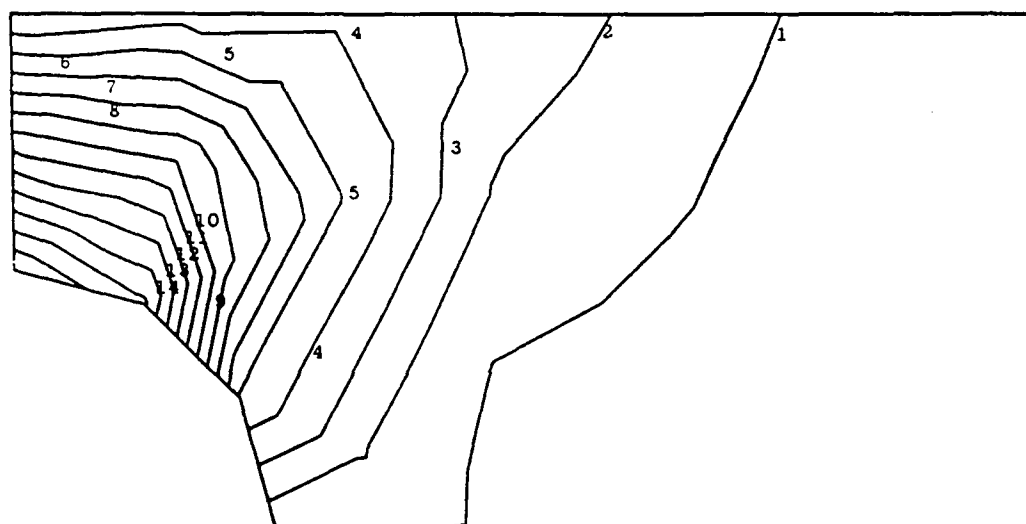
Fig. 5(b) Normalized von-Mises Stress Distribution.



Min: 0.014
Max: 0.136

0.127	15
0.120	14
0.112	13
0.105	12
0.097	11
0.089	10
0.082	9
0.074	8
0.067	7
0.059	6
0.051	5
0.044	4
0.036	3
0.029	2
0.021	1

Fig. 5(c) Adaptive Accuracy Ratio Distribution.
LSF Method.



Min: 0.08
Max: 4.86

4.56	15
4.26	14
3.96	13
3.66	12
3.36	11
3.06	10
2.77	9
2.47	8
2.17	7
1.87	6
1.57	5
1.27	4
0.97	3
0.67	2
0.37	1

Fig. 5(d) Error Ratio Distribution.
LSF Method.

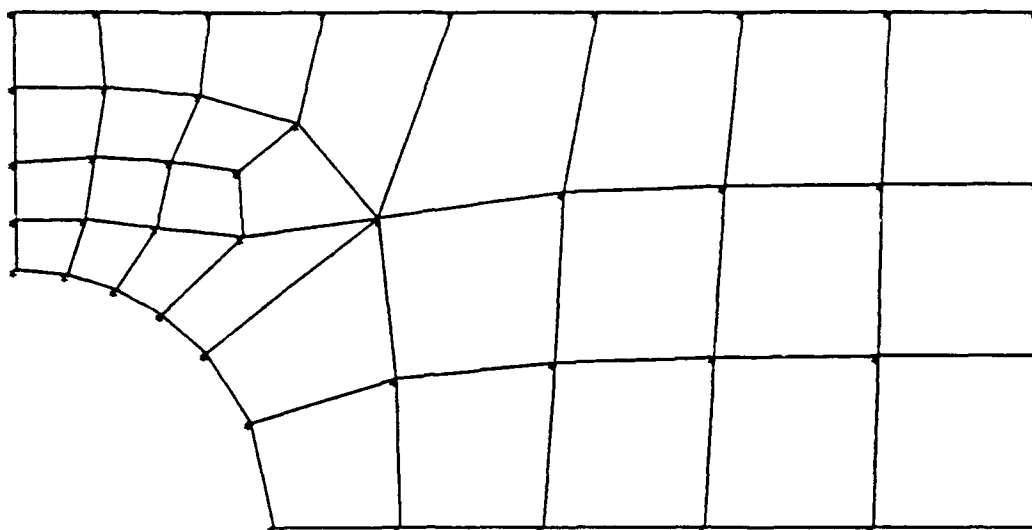


Fig. 6(a) First Remesh, LSF Method (29 Elements)

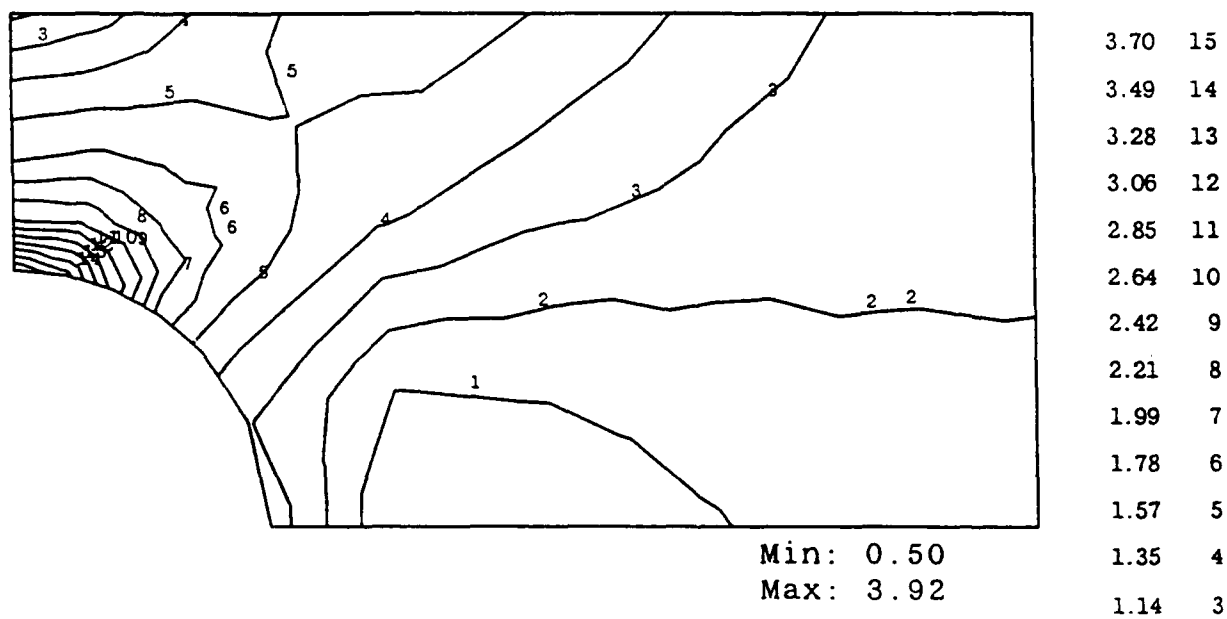
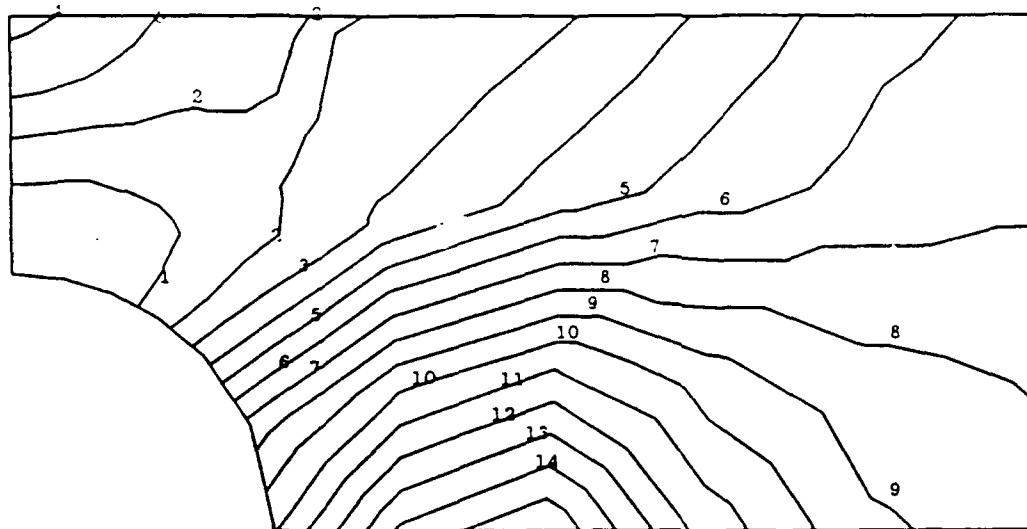


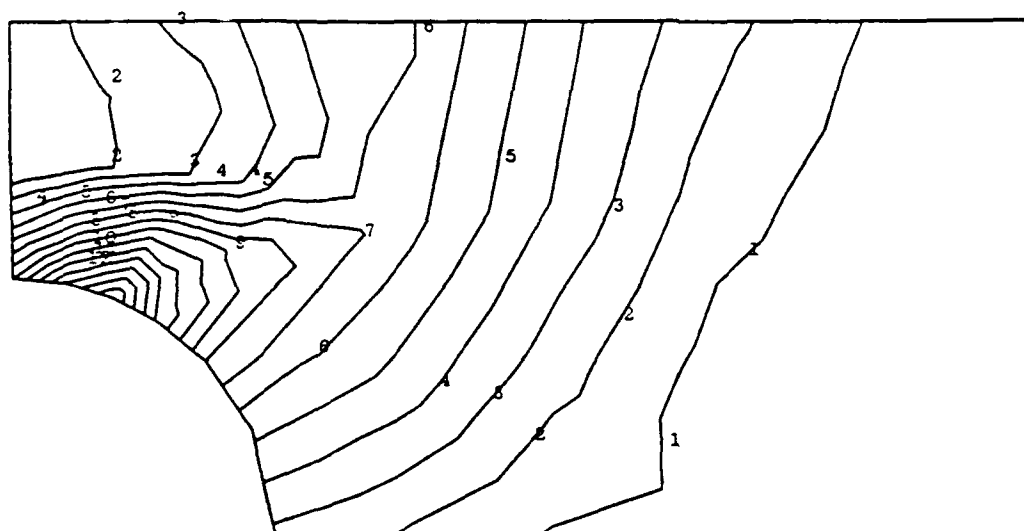
Fig. 6(b) Normalized von Mises Stress Distribution



Min:0.01
Max:0.17

Fig. 6(c) Adaptive Accuracy Ratio Distribution
LSF Method.

0.162	15
0.151	14
0.141	13
0.131	12
0.121	11
0.111	10
0.101	9
0.091	8
0.080	7
0.070	6
0.060	5
0.050	4
0.040	3
0.030	2
0.020	1



Min:0.07
Max:3.30

Fig. 6(d) Error Ratio Distribution
LSF Method.

3.10	15
2.90	14
2.70	13
2.49	12
2.29	11
2.09	10
1.89	9
1.69	8
1.49	7
1.28	6
1.08	5
0.87	4
0.67	3
0.47	2
0.27	1

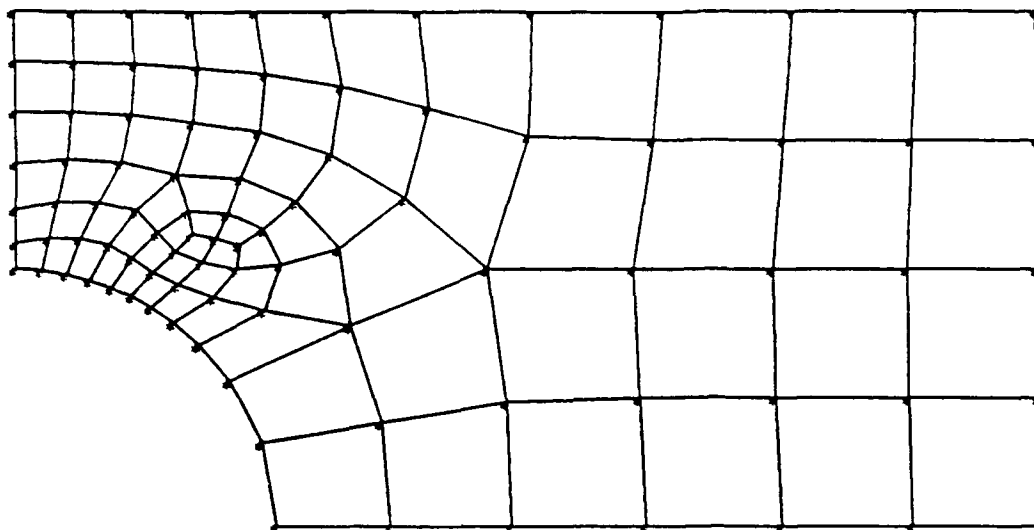
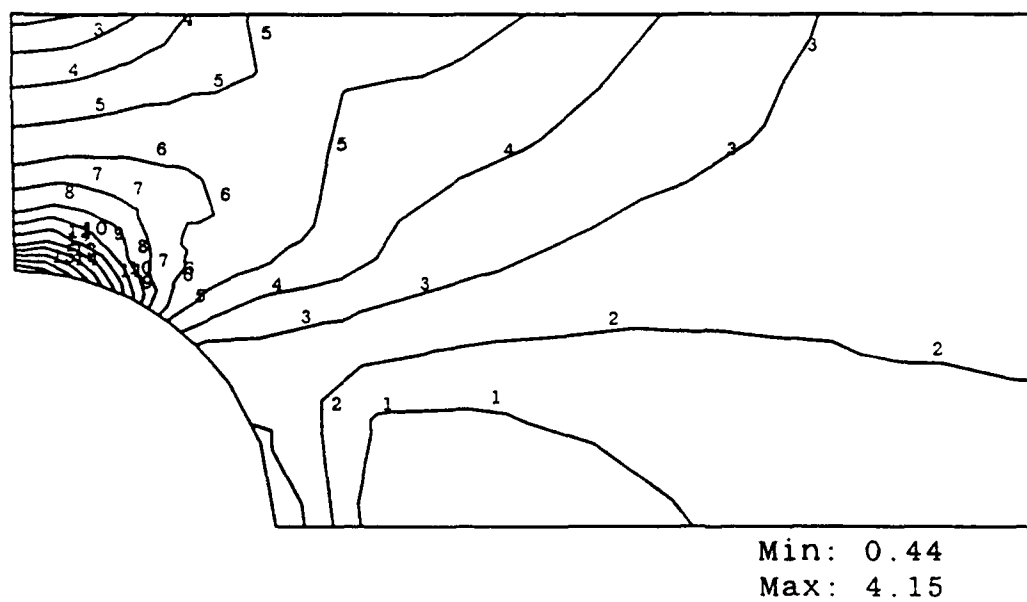
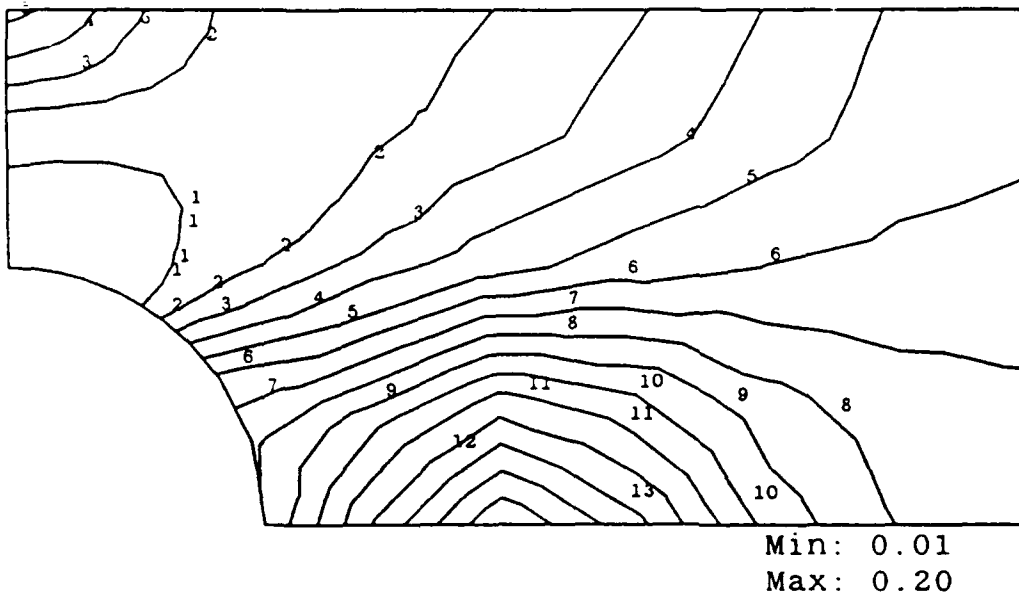


Fig. 7(a) Third (Final) Remesh, LSF Method (72 Elements)



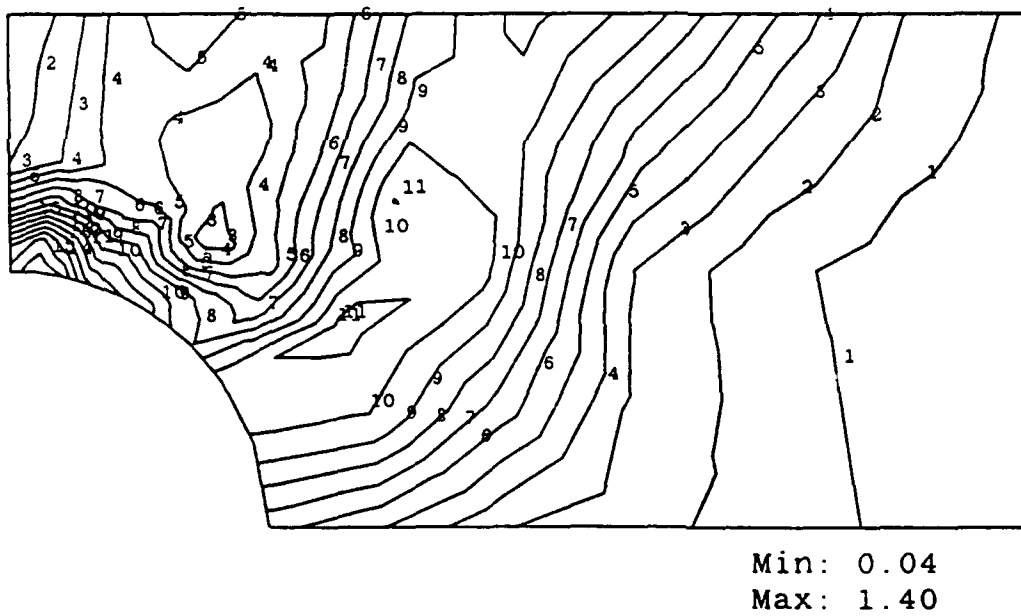
3.92	15
3.69	14
3.46	13
3.23	12
2.99	11
2.76	10
2.53	9
2.30	8
2.07	7
1.83	6
1.60	5
1.37	4
1.14	3
0.90	2
0.67	1

Fig. 7(b) Normalized von Mises Stress Distribution



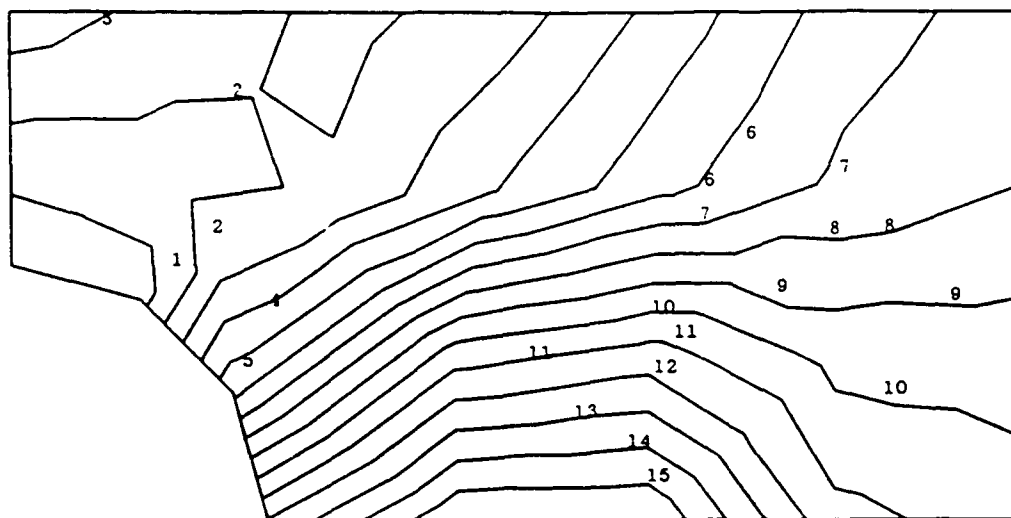
0.185	15
0.173	14
0.162	13
0.150	12
0.138	11
0.126	10
0.115	9
0.103	8
0.091	7
0.080	6
0.068	5
0.056	4
0.045	3
0.033	2
0.021	1

Fig. 7(c) Adaptive Accuracy Ratio Distribution
LSF Method.



1.31	15
1.23	14
1.14	13
1.06	12
0.97	11
0.88	10
0.80	9
0.71	8
0.63	7
0.55	6
0.46	5
0.38	4
0.29	3
0.21	2
0.12	1

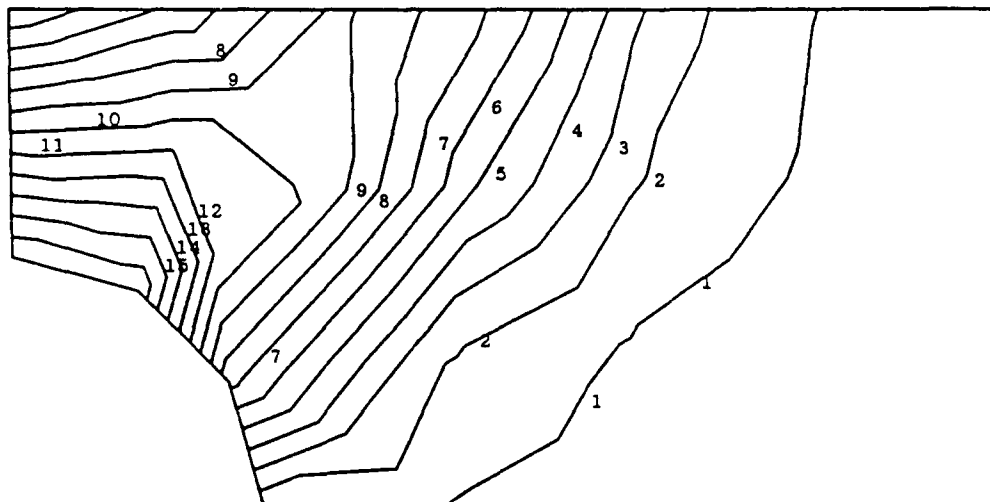
Fig. 7(d) Error Ratio Distribution.
LSF Method.



Min: 0.014
Max: 0.155

0.146	15
0.137	14
0.128	13
0.119	12
0.110	11
0.102	10
0.093	9
0.084	8
0.075	7
0.066	6
0.057	5
0.049	4
0.040	3
0.031	2
0.022	1

Fig. 8(a) Adaptive Accuracy Ratio Distribution.
SE Method.



Min: 0.06
Max: 5.98

5.61	15
5.24	14
4.87	13
4.50	12
4.13	11
3.76	10
3.39	9
3.02	8
2.65	7
2.28	6
1.91	5
1.54	4
1.17	3
0.798	2
0.428	1

Fig. 8(b) Error Ratio Distribution.
SE Method.

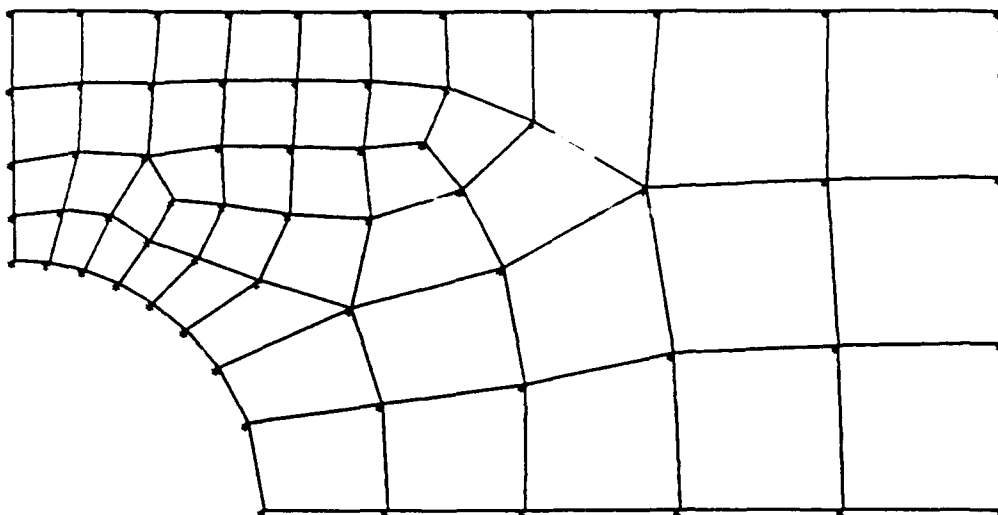
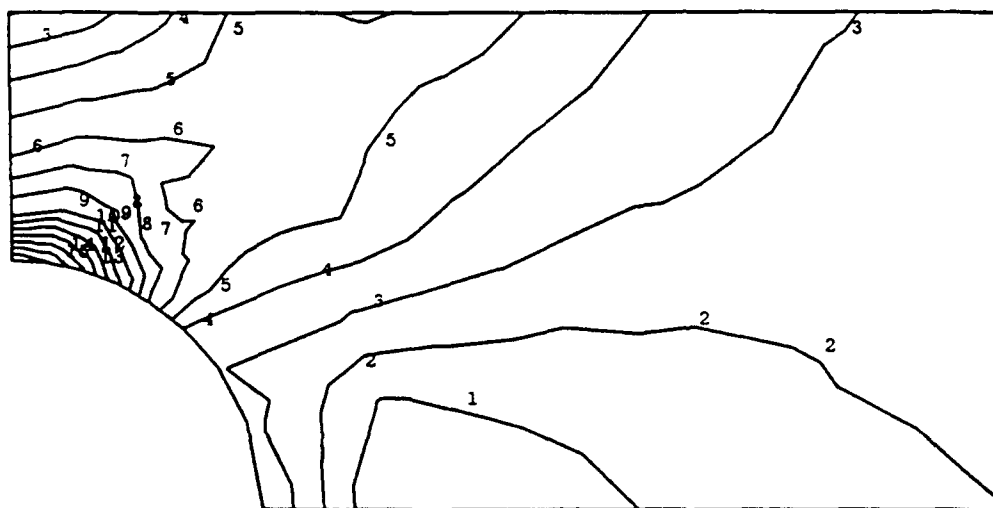


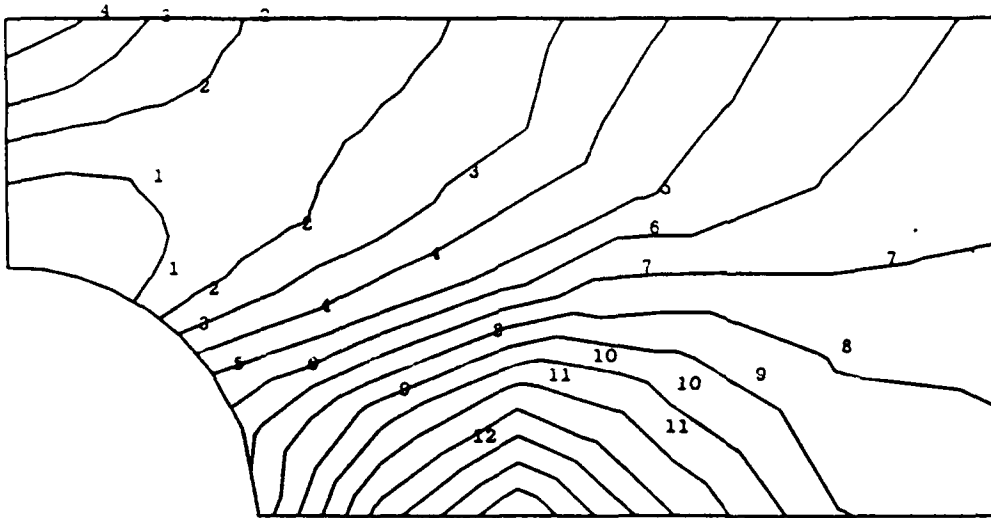
Fig.9(a) First Remesh, SE Method (45 Elements)



Min: 0.43
Max: 4.00

3.78	15
3.55	14
3.33	13
3.11	12
2.88	11
2.66	10
2.44	9
2.21	8
1.99	7
1.77	6
1.54	5
1.32	4
1.10	3
0.87	2
0.64	1

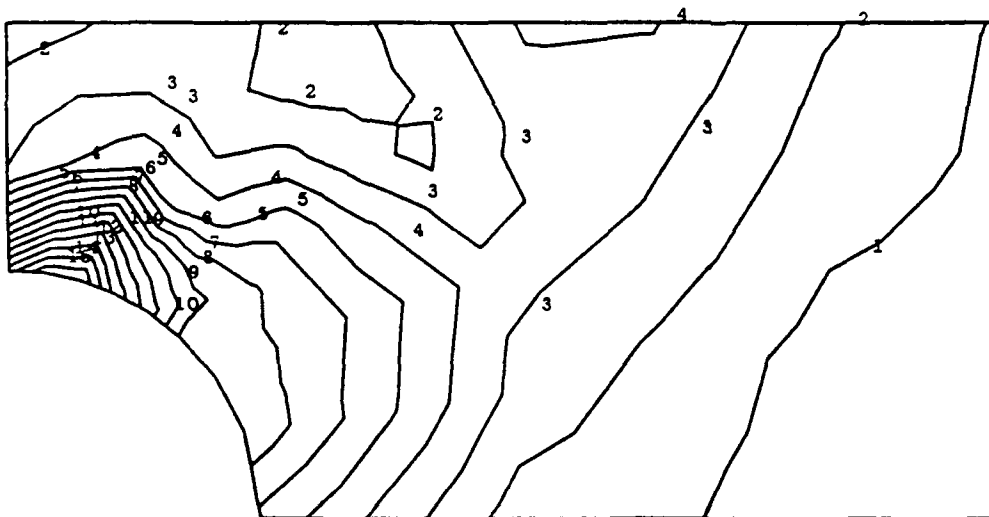
Fig.9(b) Normalized von Mises Stress Distribution



Min: 0.011
Max: 0.166

Fig. 9(c) Adaptive Accuracy Ratio Distribution
SE Method.

0.171	15
0.160	14
0.150	13
0.139	12
0.128	11
0.117	10
0.106	9
0.096	8
0.085	7
0.074	6
0.063	5
0.053	4
0.042	3
0.031	2
0.020	1



Min: 0.12
Max: 3.33

Fig. 9(d) Error Ratio Distribution
SE Method.

3.13	15
2.93	14
2.73	13
2.52	12
2.32	11
2.12	10
1.92	9
1.72	8
1.52	7
1.32	6
1.12	5
0.92	4
0.72	3
0.52	2
0.32	1

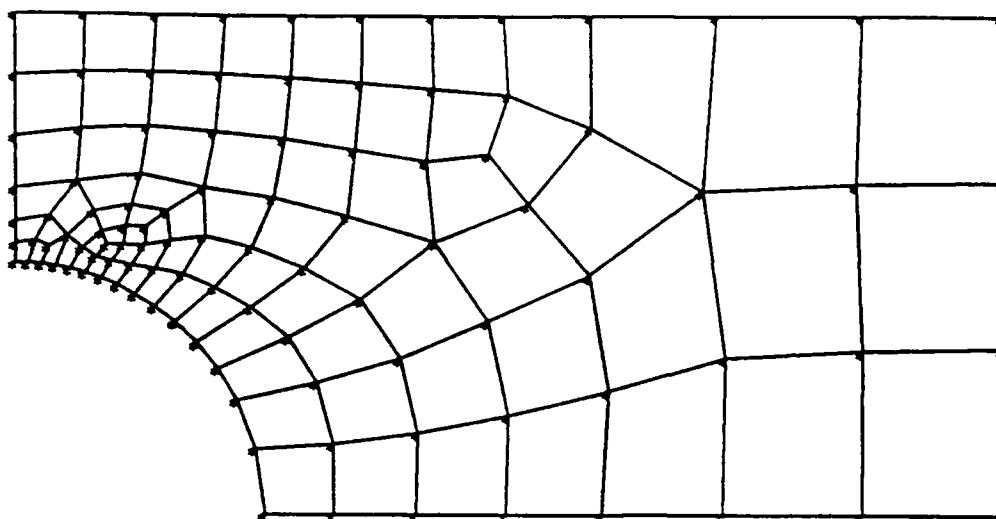
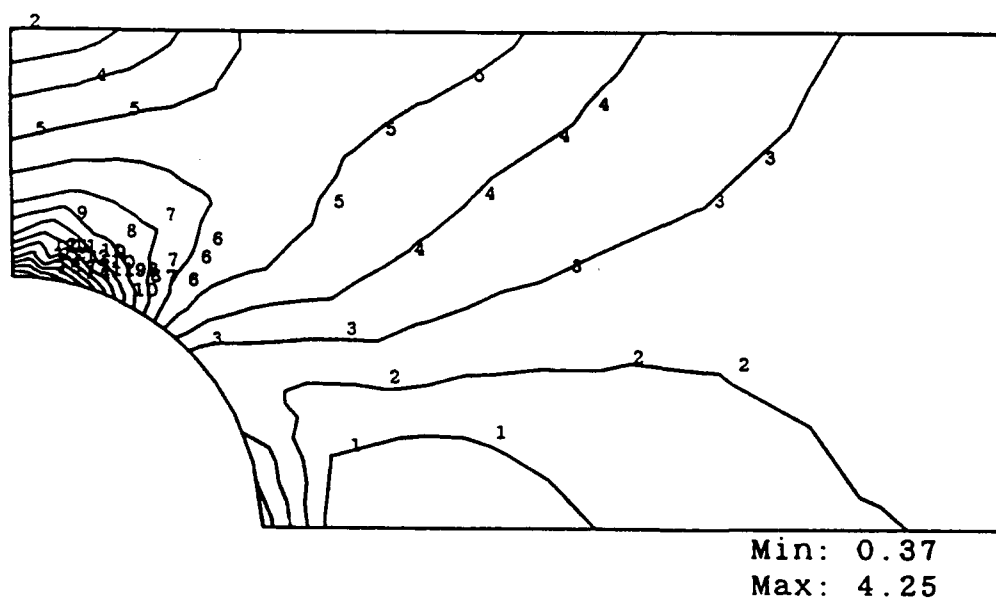
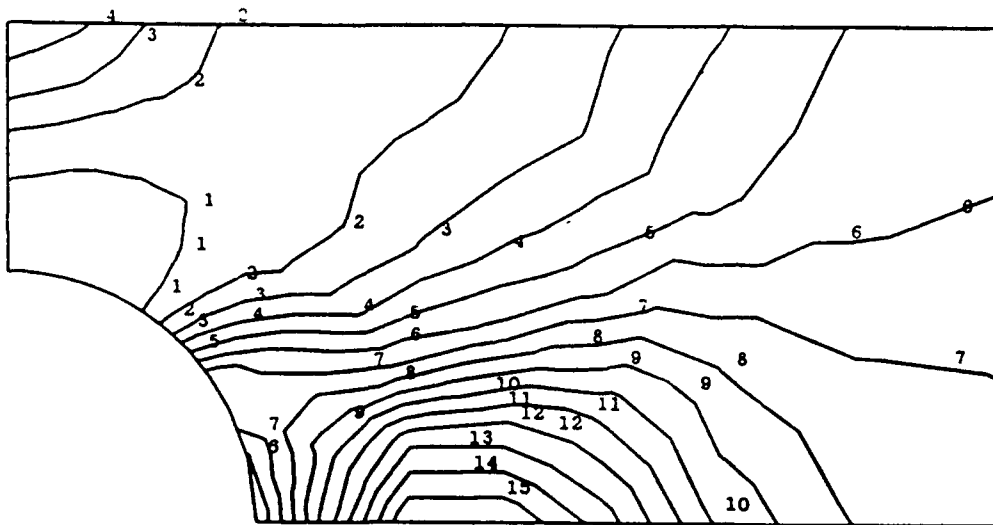


Fig. 10(a) Second Remesh, SE Method (81 Elements)



4.01	15
3.77	14
3.52	13
3.28	12
3.04	11
2.79	10
2.55	9
2.31	8
2.07	7
1.82	6
1.58	5
1.34	4
1.09	3
0.85	2
0.60	1

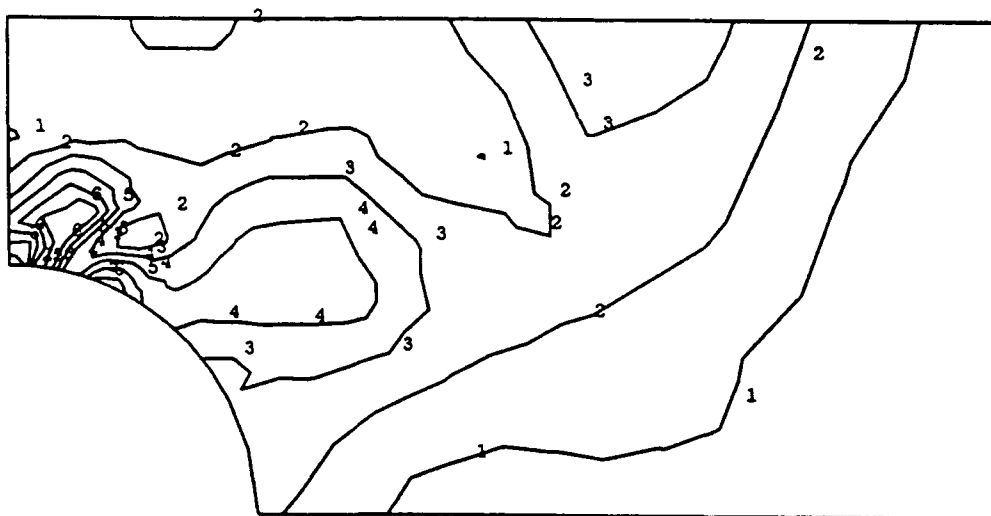
Fig. 10(b) Normalized von Mises Stress Distribution



Min: 0.01
Max: 0.20

Fig. 10(c) Adaptive Accuracy Ratio Distribution
SE Method.

0.18	15
0.17	14
0.16	13
0.15	12
0.14	11
0.12	10
0.11	9
0.10	8
0.09	7
0.08	6
0.06	5
0.05	4
0.04	3
0.03	2
0.02	1



Min: 0.11
Max: 1.60

Fig. 10(d) Error Ratio Distribution, SE Method

1.39	6
1.18	5
0.96	4
0.75	3
0.53	2
0.32	1

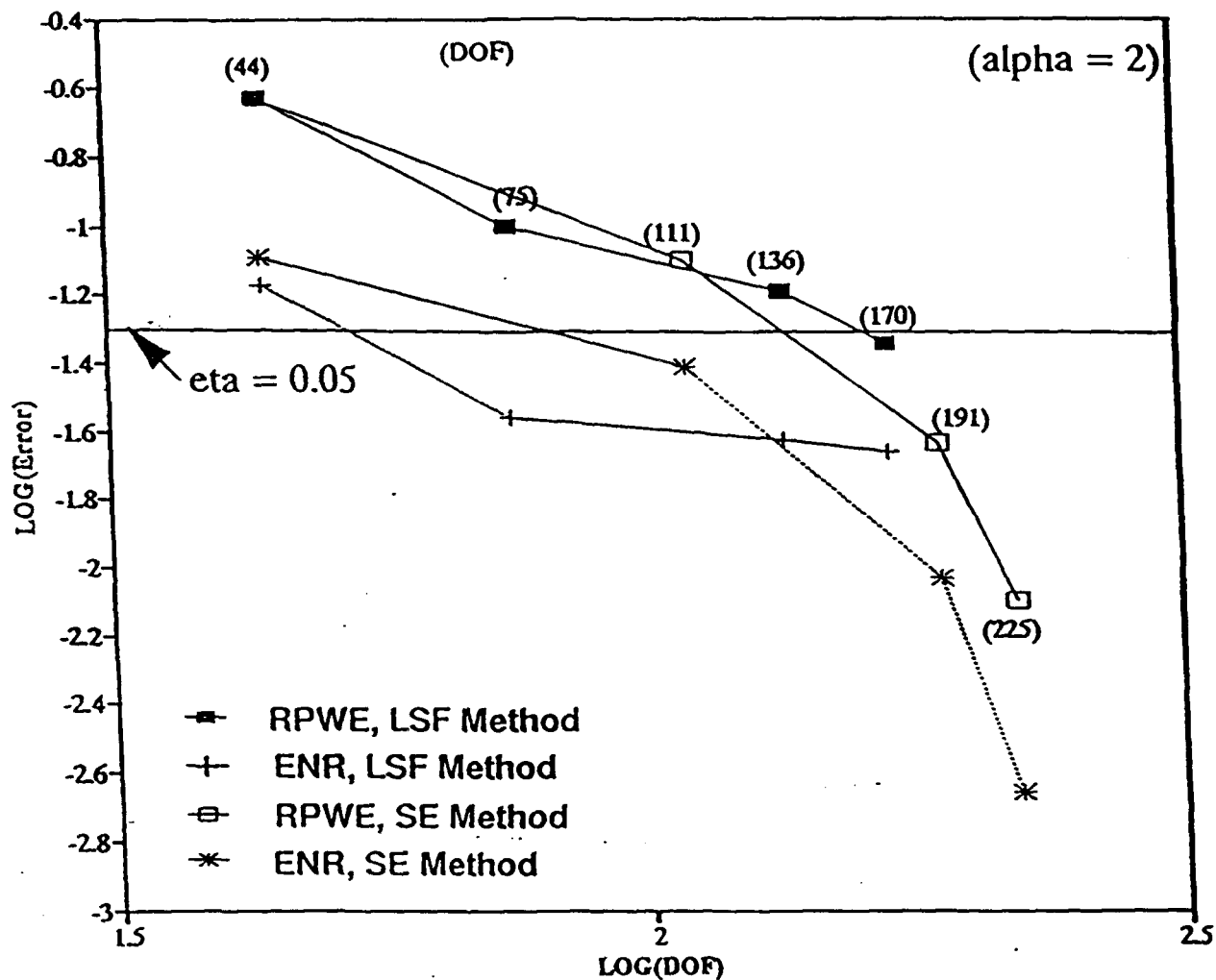


Fig. 11(a): Convergence Comparison of LSF and SE Methods

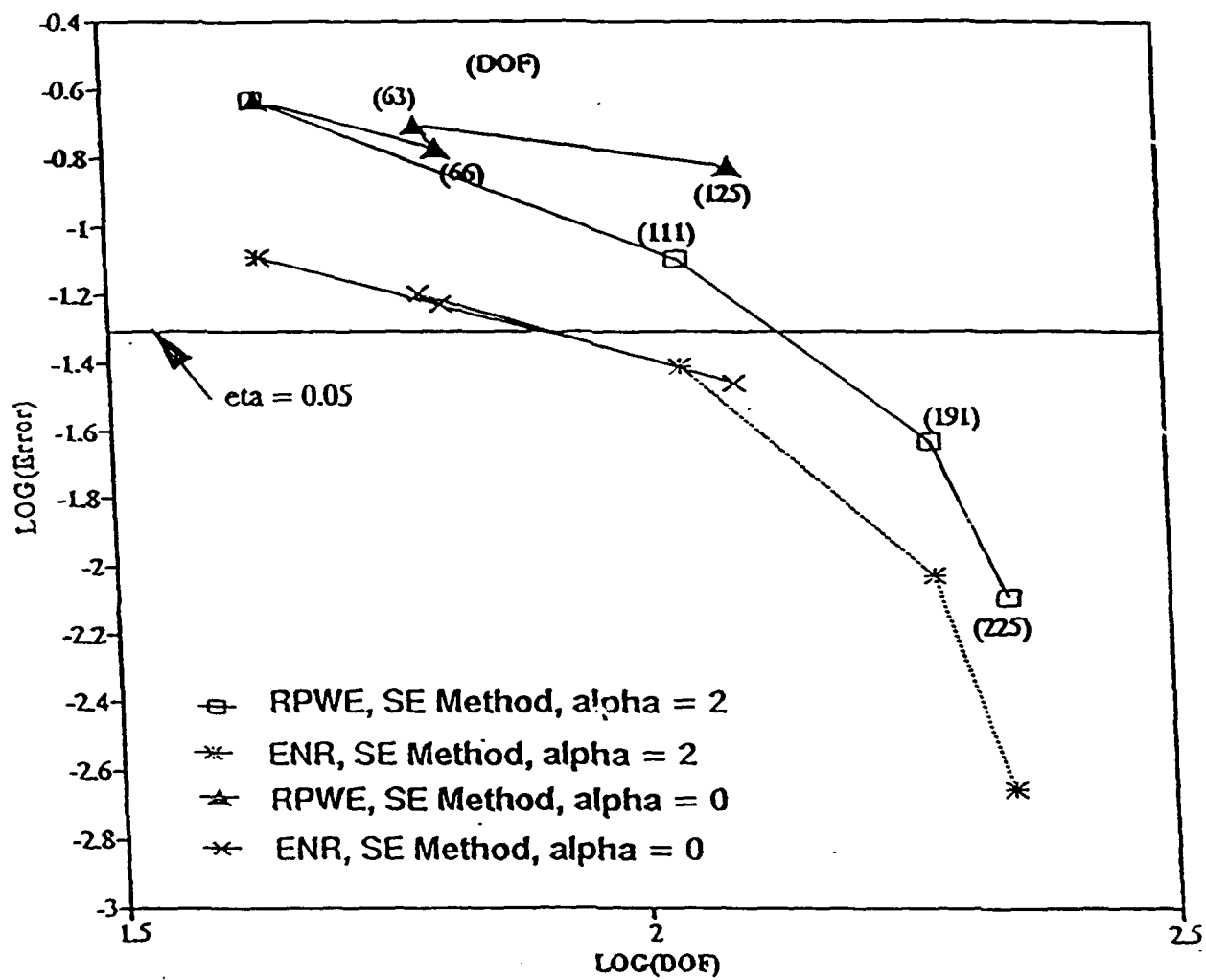


Fig. 11(b): Convergence Comparison of Nonadaptive vs. Adaptive Accuracy

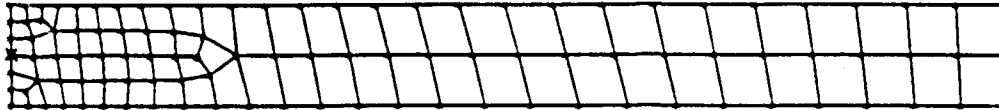


Fig.13(a) First Remesh, SE Method (74 Elements)

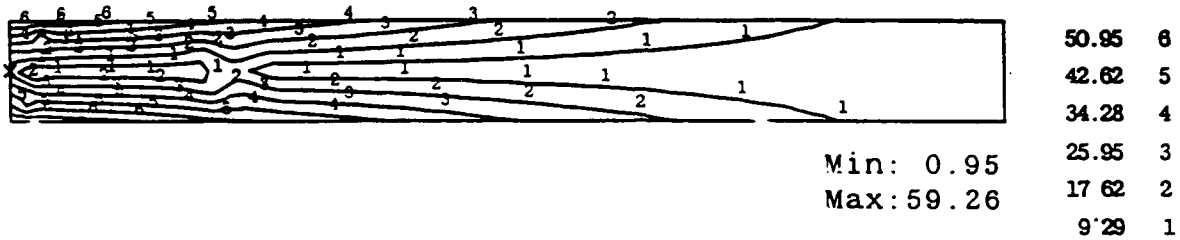


Fig.13(b) Normalized von Mises Stress Distribution

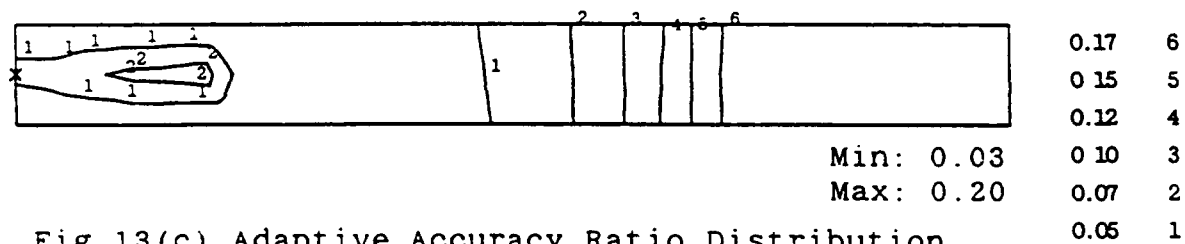


Fig.13(c) Adaptive Accuracy Ratio Distribution

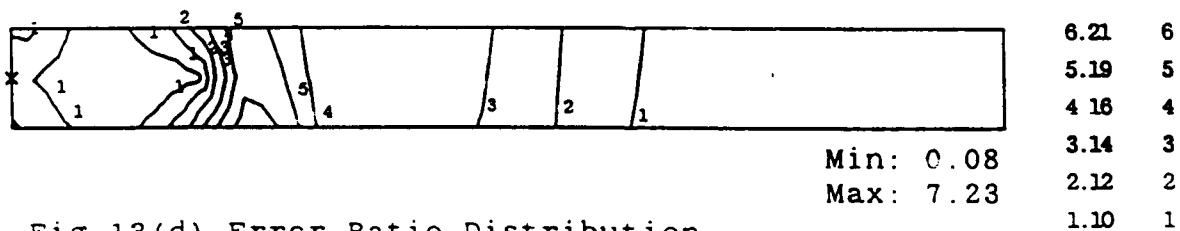


Fig.13(d) Error Ratio Distribution

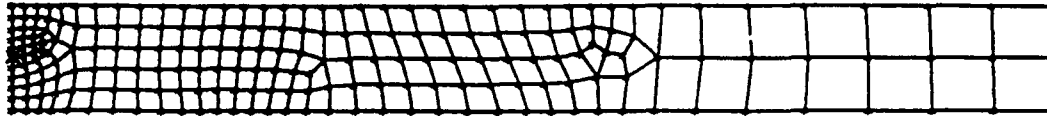


Fig.14(a) Final (Third) Remesh, SE Method (168 Elements)

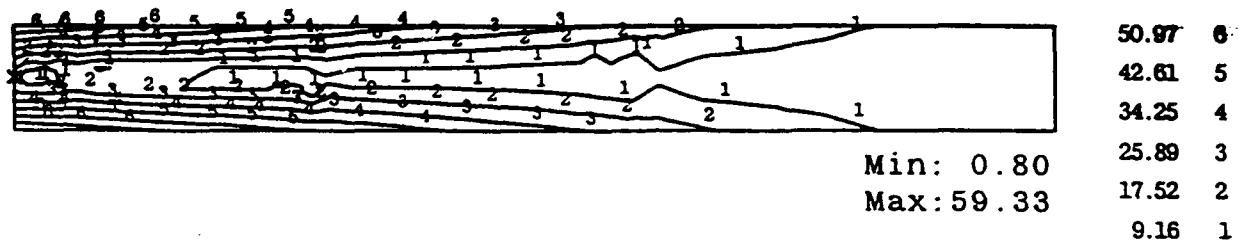


Fig.14(b) Normalized von Mises Stress Distribution

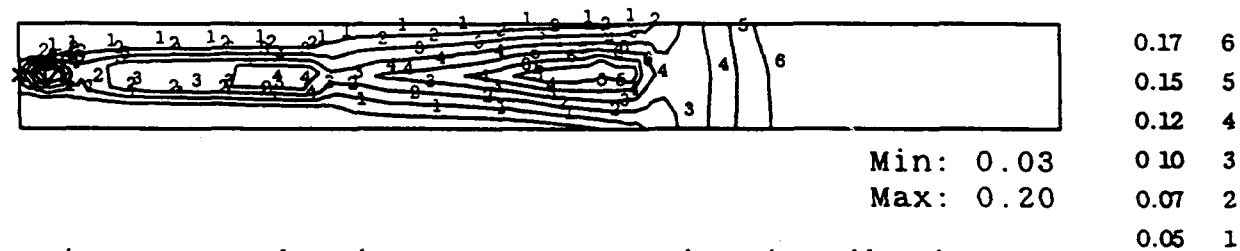


Fig.14(c) Adaptive Accuracy Ratio Distribution

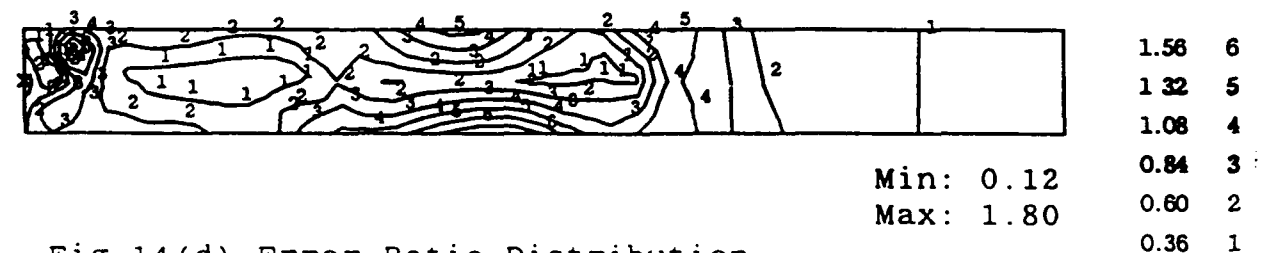


Fig.14(d) Error Ratio Distribution

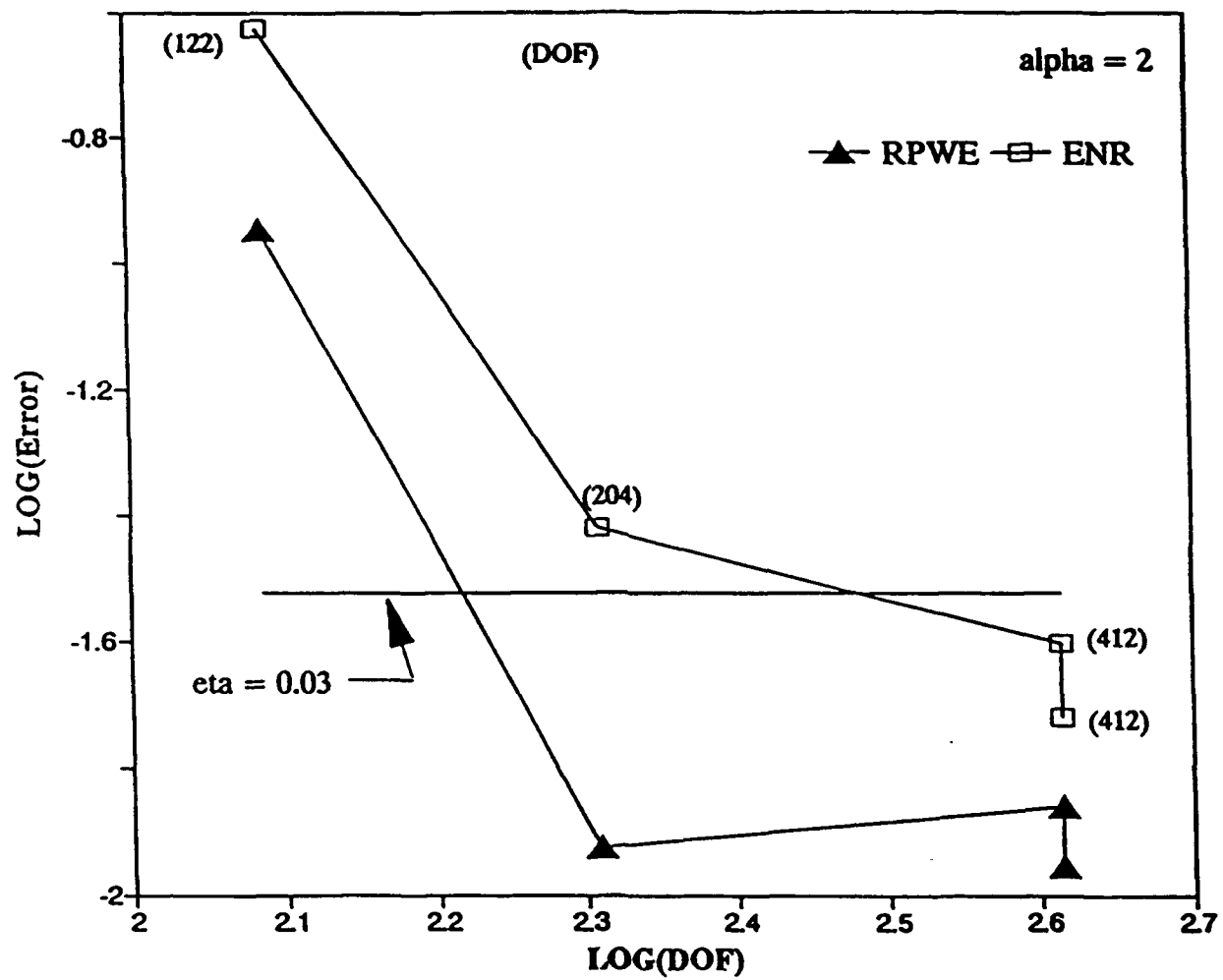


Fig. 15: Convergence of the Error for the Beam Example

Final Report

**Ionospherically-Induced Phase Distortion Across Wide-Aperture
HF Phased Arrays**

Henry F. Helmken

and

Ronald Chilluffo

**Florida Atlantic University
Boca Raton, Florida**

December 31, 1991

Abstract: This report presents some measurements of HF received signal phase characteristics. Phase distortion across a large aperture array is examined and origins and initial experimental measures of these phase errors are examined. The effects of propagation mode separation are detailed and an experimental approach providing a statistical description of phase distortion is formulated. Experimental measurements were made using both E-W and N-S linear array baselines for a one-hop, 2400 km., near geomagnetic north-south path between Ava, NY and Boca Raton, FL. Initial unmodulated CW characterization is described and test results summarized. A second set of experiments involving a 127 length PN code, was used to study mode separation. An extensive bibliography highlighting past and current research in ionospheric path characterization is presented.

1.0 Introduction

For an ideal phased array radar, angular resolution is directly proportional to receiver array aperture [1] and the phase coherence across the aperture. Unlike Line-Of-Sight (LOS) systems, Over-The-Horizon (OTH) HF radars [2] rely on refraction through the ionosphere which can bring an attendant degradation of phase coherence. As apertures are increased, a "coherence-limited" phenomenon is observed which limits angular resolution and causes beamshape spoilage [3]. This phase distortion is quantified by root-mean-square (rms) phase error and correlation length, i.e. length where spatial cross correlation coefficient reduces by $1/e$ [4]. Although research in this area is not extensive, some results indicate correlation lengths on the order of 1 kilometer, corresponding to 10-100 wavelengths in the 3-30 MHz HF band [5]. Since larger apertures are desirable for improved angular resolution, additional knowledge of phase coherence across wide-aperture HF phased arrays is needed. With this information, system performance can be analyzed and optimum operational parameters (eg. frequency, modulation) can be selected.

The determination of phase error and correlation length is the subject of this research. The experimental characterization required special signal processing techniques to assure accuracy of wide-aperture phase distortion characteristics. A statistical description is selected over a theoretical or empirical approach because results can be directly entered into the radar performance evaluations [6]. The phase coherence statistics involve correlation measures of spatial, temporal, and for wide bandwidth, frequency coherence. Beamshape spoilage and pointing error statistics directly follow from this phase coherence statistical description [7]

A complete description would require analysis of relevant ionospheric parameters including hourly solar flux/geomagnetic activity vs. time-of-day, day-of-month, month-of-year, and year-of-11-year solar cycle. These effects combine and can be viewed as a continuously changing, distorted wavefront impinging on the

wide-aperture array [8].

The key obstacle to accurate experimental characterization lies in ionospheric propagation mode separation [9]. Although a single, linearly polarized wavefront may be launched from the transmit site, the signal arriving at the receive site will contain components of orthogonal polarization [10], differing path lengths [10], Doppler shifts [11], and amplitude fading [12].

Since the focus of this research is phase coherence, path length separation is addressed. This necessitates employing special mode separation signal processing techniques and constitutes the challenge in this investigation. The other identified phenomena are not encompassed. Path length separation signal processing techniques are similar to ranging techniques utilized in radar and navigation application [13]. Innovative signal processing techniques were required to enhance mode-separation accuracy.

2.0 Background

Many years of extensive research have provided much knowledge of both ionospheric propagation and the underlying physical processes responsible for these characteristics [14], [15]. Ionospheric propagation amplitude characteristics have received the bulk of the emphasis since point-to-point AM communications have been the primary application. Phase characteristics have received less attention due to their limited applications, viz. OTH radar. These phase characteristics are quantified by phase error and spatial correlation length. Phase error is the phase difference between a signal that propagates through the ionosphere and a signal that propagates through an equi-distant ideal transmission medium. In general for an N element array, these phase errors relative to some ideal reference are of the form shown in Figure 1.1. In the absence of this ideal reference, the signal from element #1 can be used as reference. In this case, Figure 1.2 depicts this more-readily measurable phase differences between progressively separated array elements. Correlation length is determined from the cross correlation function of these phase

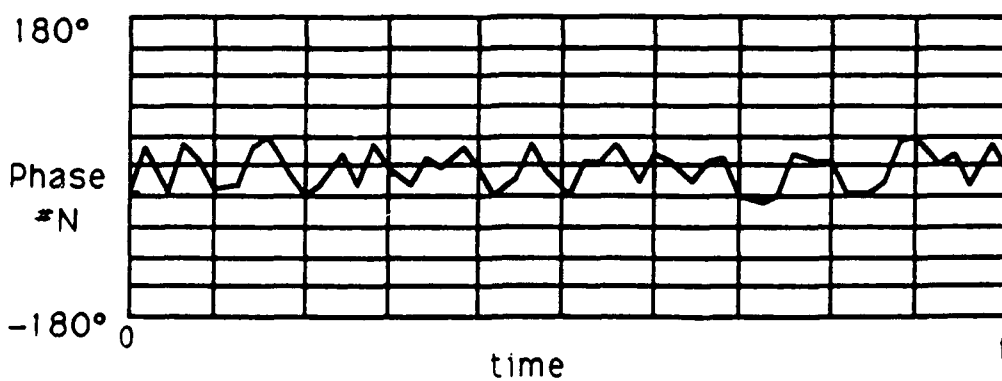
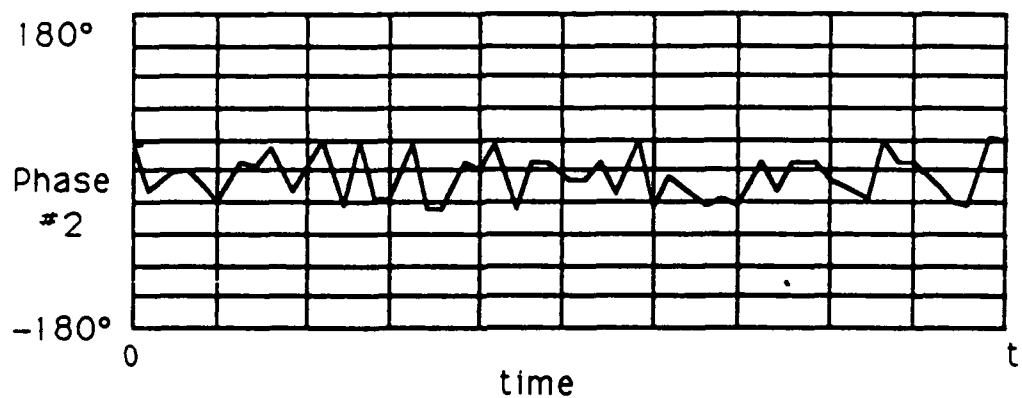
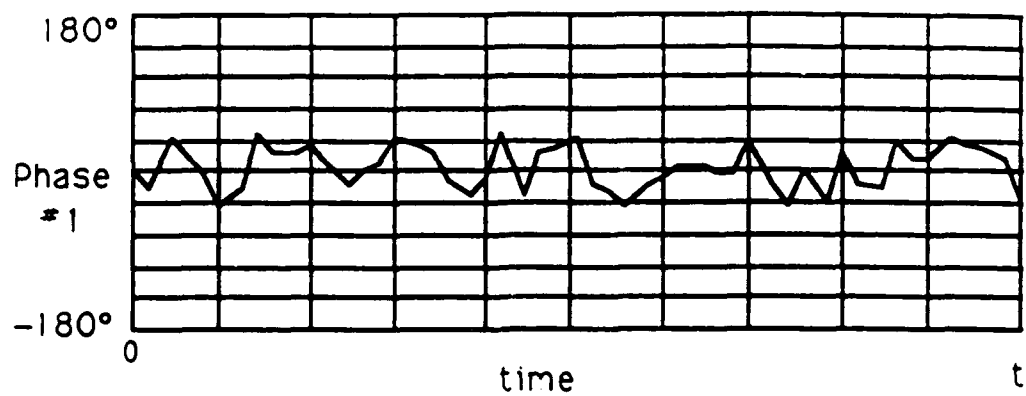


Figure 1.1 Typical Element Phases Relative To A Reference Signal That Propagated Through An Ideal Medium

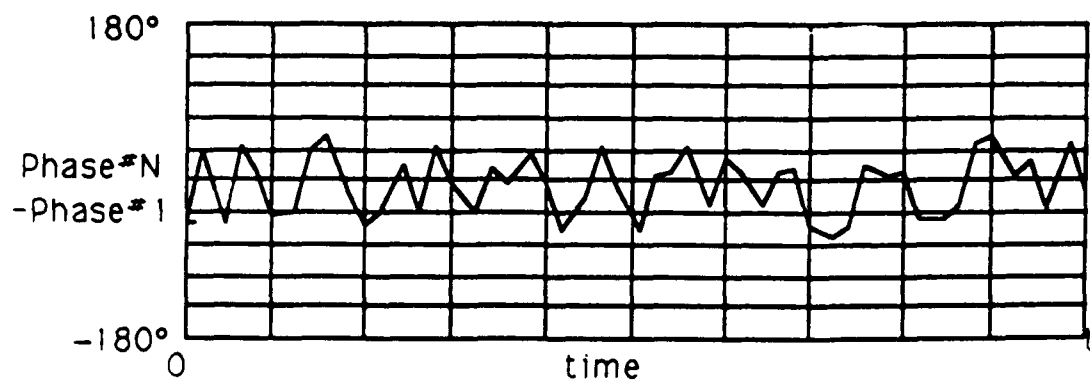
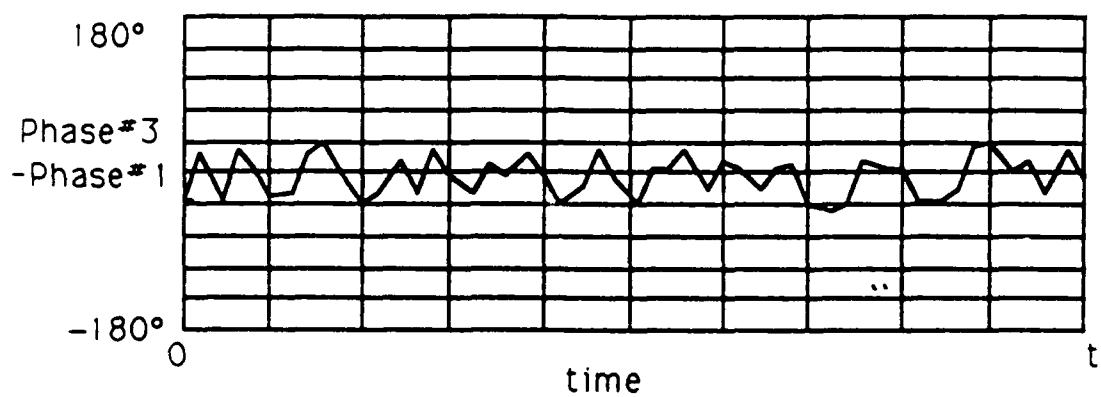
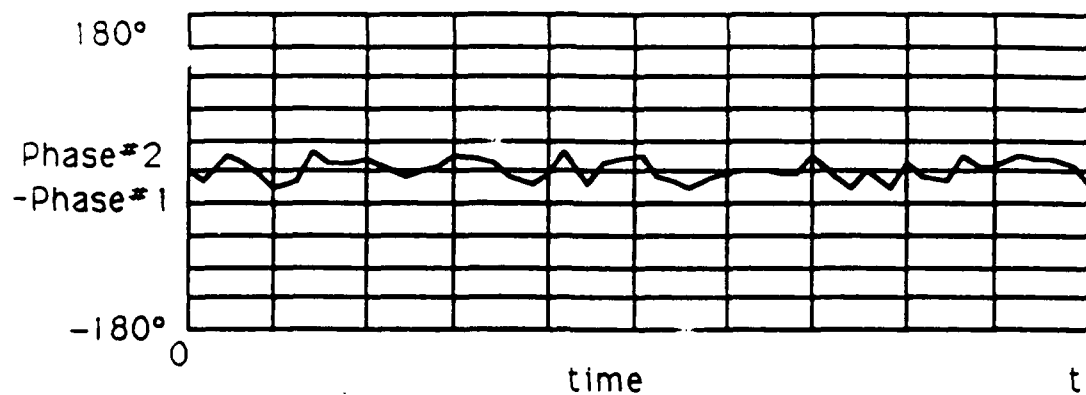


Figure 1.2 Typical Element Phases Relative To Element #1

error results. It is noted that phase difference in Figure 1.2 varies from zero to twice the phase error of Figure 1.1 as the element separation increases from zero to lengths much greater than the correlation length, i.e. when the two elements become uncorrelated.

1.1 Origin of Phase Error

The ionosphere consists of stratified layers (75-400km.) of ions (plasma) which are generated by solar radiation and subsequently influenced by the geomagnetic field. Refractive effects of these ionic layers cause HF electromagnetic ray path bending. The predominate layer is the high-altitude (350km.) F-layer (night) which separates during daytime into two layers, F1 and F2 [16]. An intermediate-altitude (100km.) E-layer also provides additional propagation paths. During daytime, a low altitude (75km.) D-layer is produced which attenuates lower frequency waves (≤ 10 MHz). This layer rapidly dissipates after sunset.

Solar flux (radiation) affects the number of ions created, while terrestrial geomagnetic activity affects the disturbances and motions of these ions. Solar radiation is related to sunspot number. It is characterized by daily radio solar flux measurements at Ottawa, Canada (2695 MHz). Geomagnetic activity is quantified by K and A indices and is monitored at multiple observatories around the world. Broadcasts of solar flux and geomagnetic activity are available on 5, 10, 15, and 20 MHz from WWV in Boulder, Colorado. This information is also published periodically. Any Sudden Ionospheric Disturbances (SID's) are indicated by an associated sun-earth 8 minute radiation transit time and 40 hour particle transit time. Message information is also available from the Space Environment Laboratory.

Since ionospheric propagation is so dependent on solar activity, there are cyclical effects related to time-of-day (diurnal effects), day-of-month (27 day solar rotation period), month-of-year (seasonal effects), and year-of-11-year solar cycle (solar phenomenon). Since the latest Cycle 22 peaked in 1990, this research encompasses a snapshot immediately following this peak.

1.2 Ionospheric Status

Ionospheric propagation frequency characteristics are measured at numerous sites around the world [17]. Bottom-side sounding (ionograms, sweep the ionosphere through the HF band and measure characteristics such as critical frequency, Maximum-Usable-Frequency (MUF) and Lowest-Usable-Frequency (LUF), from which the Optimum-Working-Frequency (OWF) or Optimum Traffic Frequency (OTF) can be derived [18]. Top-side sounding had been conducted from satellites [19] but no active program currently exists. These experimental results have aided in the development of ionospheric frequency propagation models [20].

Ionospheric physical composition has also been researched extensively yielding three distinctive regions: polar, mid-latitude, and equatorial [21]. The polar regions exhibit distinctive characteristics due to the accumulation of solar particles around the poles, while the equatorial region with its abundant solar radiation exhibits distinctive characteristics, eg. enhanced propagation along trans-equatorial (TE) north-south paths. Also, gray-line propagation along the Earth's terminator exhibits distinctive characteristics due to the transient behavior of ionic concentrations in the upper and lower layers. The gray-line is the band of twilight that separates the white (day) and black (night) regions of the Earth [22].

1.3 Multipath

Direction of propagation relative to the geomagnetic field (B_0) gives rise to differing propagation characteristics. Faraday rotation is the rotation of polarization due to the Lorentz force acting on electrons as a wave propagates in the presence of B_0 , i.e. $q(\mathbf{v} \times \mathbf{B}_0)$ [23]. For propagation parallel to B_0 , a linearly polarized wave can be decomposed into a right-hand circular (RHC) and left-hand circular (LHC) components. Each experiences differing phase velocity and attenuation parameters resulting in a net linear polarization rotation.

For propagation perpendicular to B_0 , a linearly polarized

wave consists of a component parallel to B_0 (ordinary component and unaffected by B_0) and a component perpendicular to B_0 (extraordinary component and rotated by B_0). A given ray path will experience a complicated combination of these effects; the resultant effect is a signal separation at the receive site. The IEEE Standard Definition of Terms for Radio-Wave Propagation states the ordinary wave component "deviates the least relative to a wave in the absence of B_0 ", while the extraordinary wave component has its "electric vector in the opposite sense to that of the ordinary wave component" [24].

1.4 Current Research

This research investigates propagation along a North-South geomagnetic field-aligned path. Related phase characteristic work has been done in the area of radio astronomy along very long baselines [25]. Scintillation effects are similar to what is encountered in this investigation, but important differences do exist. This research addresses oblique (near-tangential) as opposed to normal (near-perpendicular) incidence and a deterministic as opposed to noise-like signal characteristics.

Steinberg [26] investigates methods of increasing HF radar azimuthal resolution. Crane [27] reports theory and experimental results of ionospheric scintillation effects on angle-of-arrival and Doppler fluctuations. Humphrey [28] examines multi-path phase fluctuation over a Panama-New York one-hop path. Kieburzt [29] analyzes angle-of-arrival measurements by phase differences in a two-element interferometer. Barrick [30] reports techniques and results of wide-aperture phase distortion measurements between California-Arkansas.

Signal processing research related to ionospheric characterization is quite sparse. Some work in support of frequency characterization (sounding) has been reported by Coll, et.al. [31]. Although FM/CW is a popular ranging technique, other modulations can provide improvements in characterization accuracy. Application of spread-spectrum Pseudo-Noise (PN) and complementary code (CC) signal processing techniques offers great promise for

expanding the body of knowledge of ionospheric propagation [32-33].

2.0 CW Characterization

The basic aim of the CW characterization is to receive an ionospherically refracted signal at multiple elements of two perpendicular linear arrays and deduce azimuth and elevation angle-of-arrival from these measurements. The mean of the angle-of-arrival represents the direction of the incoming wavefront and variations of angle-of-arrival represents ionospherically-induced distortion or wander. Since long baselines and multi-wavelength element spacings are desired, angle-of-arrival ambiguities due to grating lobes must be resolved by using techniques such as multiple-frequencies, multiple-spacings, etc. In this context, an ambiguity number n denotes the integral number of wavelengths in some electrical length geometric parameter.

Due to its simplicity, unmodulated CW tests were selected for the initial characterization. Since CW does not incorporate any mode separation signal processing, the higher end of the HF band (> 10 MHz) was selected since ordinary-extraordinary wave mode separation is greatly reduced in this frequency range.

A one-hop F-layer near field-aligned path exists nearly always in the daytime between Ava, NY and FAU at Boca Raton FL. This 2400 km. path has its midpoint over North Carolina and the ionospheric interaction is confined to a length of approximately 100 km. Within this F-layer one-hop path, typical path length differences are expected to be on the order of tens of microseconds. Multipath involving other layers, path length differences are expected to be on the order of hundreds of microseconds. The deviation from exact geomagnetic field alignment is 10° , i.e. the difference between the FAU 2.5° and Ava, NY 12.5° magnetic declinations. A site's magnetic declination is the deviation between magnetic North and true North. Relative to FAU, map data indicate angle-of-arrival to Ava, NY is 12.5° azimuth (NE) and 6° - 12° elevation (corresponding to virtual heights of 200-400 km.).

2.1 Experimental Set-up

E-W and N-S baselines were laid out as shown in Figure 2.1. Element locations are labeled as indicated. Existent surveyed boundaries such as roads and runways were used for coarse positioning (± 10 cm.). Although the N-S baseline could not be visually sighted for straightness, the E-W baseline was in open field and binocular sighting of straightness yielded a ± 2 cm. fine position adjustment. Figure 2.2 depicts the E-W and N-S baseline system configurations.

The antenna elements selected were 3 m. verticals which are approximately quarter wavelength at 20 MHz. These were constructed with 1.9 cm. (3/4 in.) aluminum conduit threaded on both ends.

The lower end is screwed into a dielectric (PVC) coupling and a 0.75 m. ground rod. When 10 MHz operation is desired, a second 3 m. section is added providing an approximate quarter wavelength at 10 MHz. This ground system provided satisfactory results due to the high ground conductivity in this geographic area and the receive-only test program. Receive-only minimizes the importance of antenna efficiency since, in the HF band, atmospheric noise dominates receiver thermal noise and receive signal-to-noise ratio is not degraded by low antenna efficiency or receiver noise figure. Also, the simplicity of this ground system enabled quick assembly and disassembly.

The transmission lines were Government-furnished RG-214 in 145 m. spools. HF network analyzer electrical phase length measurements were performed and lengths trimmed to within $\pm 0.5^\circ$ at 20 MHz. Although exact electrical length equalization is desired, small deviations due to cable spooling, unspooling, flexing, etc. are calibrated out by a "central antenna" calibration procedure described below.

The receiver/controller equipment consisted of two Government-furnished Racal R-2174 HF receivers, Hewlett-Packard 35A Phase Meter, Tektronix 7D20 Programmable Digitizer, Tektronix 2214 Digital Oscilloscope, and a Macintosh computer for data storage. In addition, an RF transfer switching function is

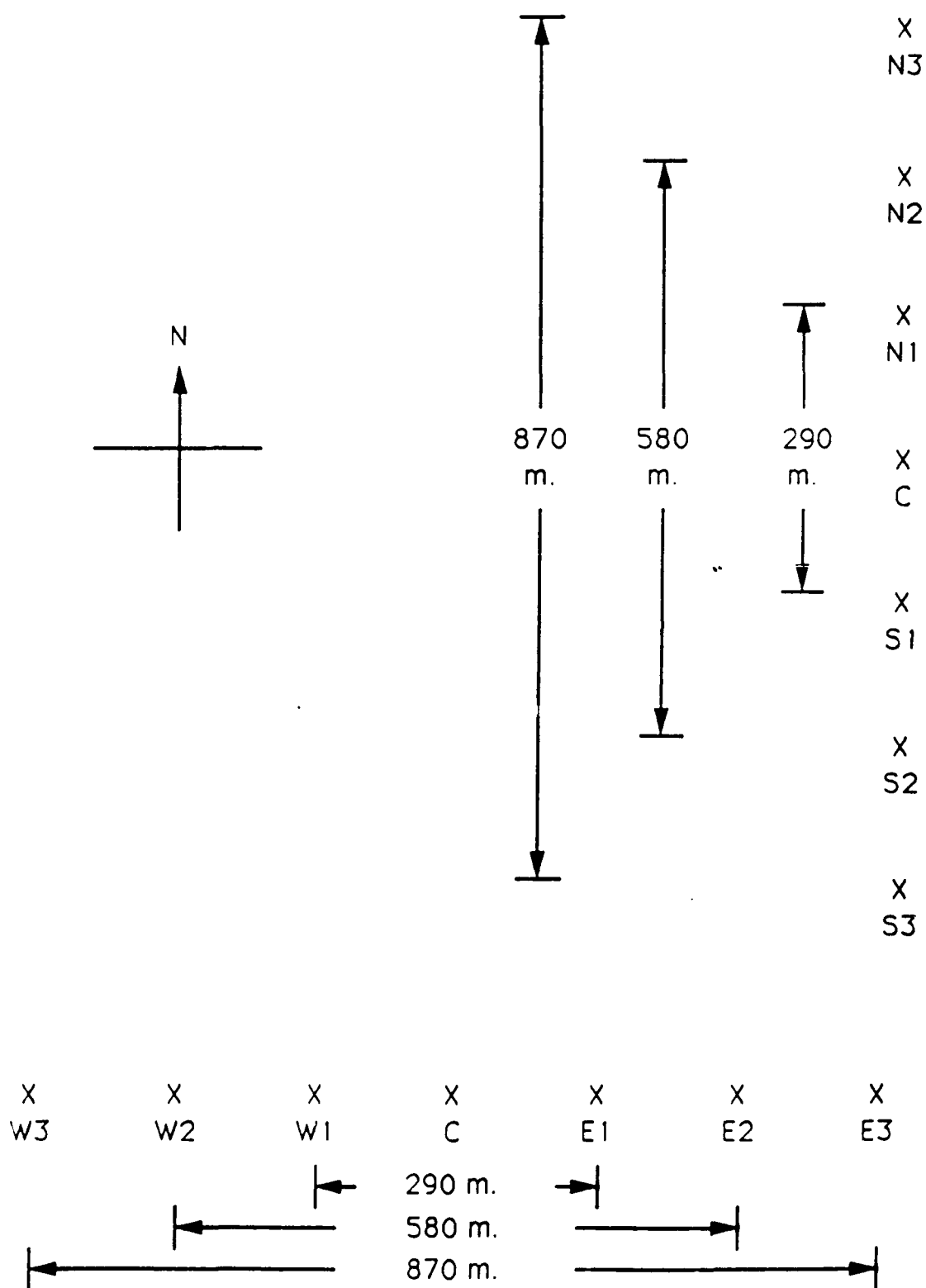
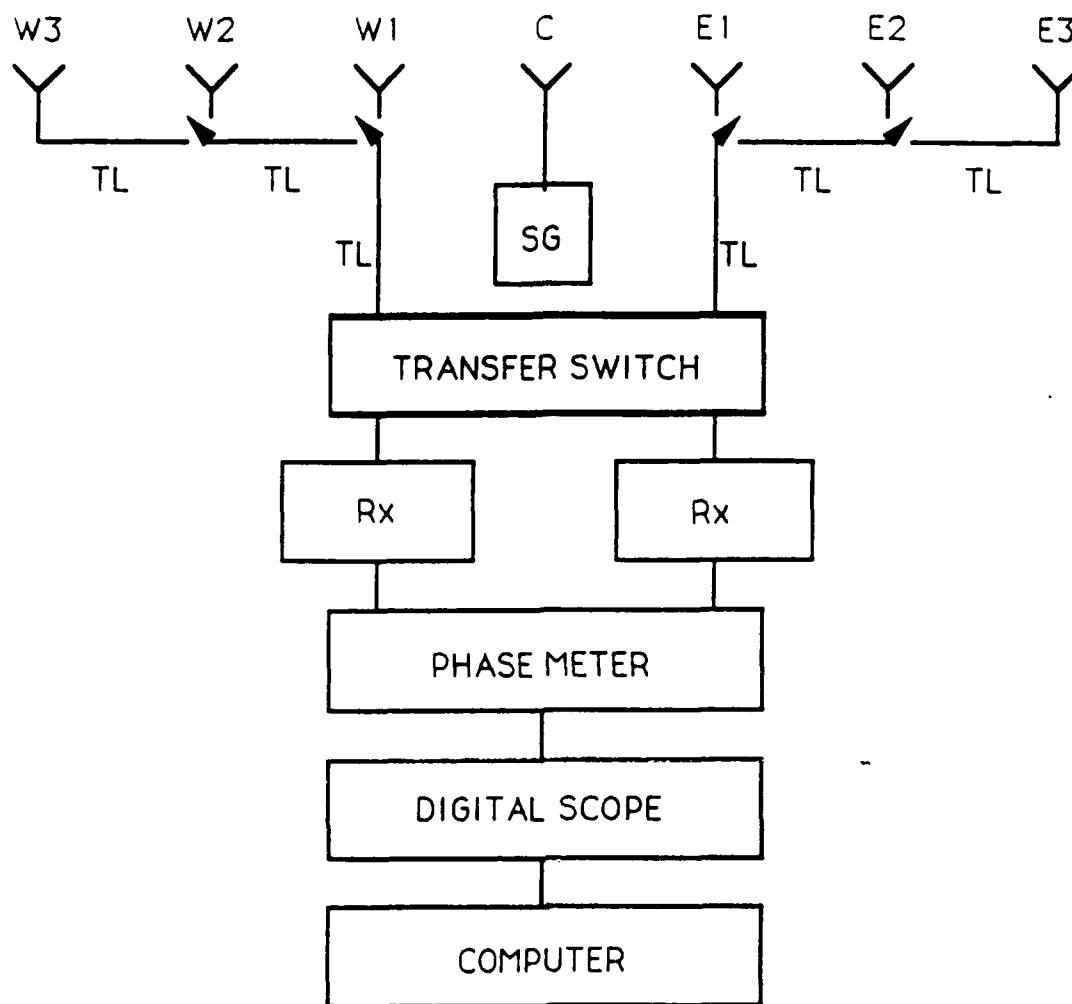


Figure 2.1 East-West and North-South Baselines



TL: RG-214 (145 m. Section)

SG: Signal Generator (Calibration)

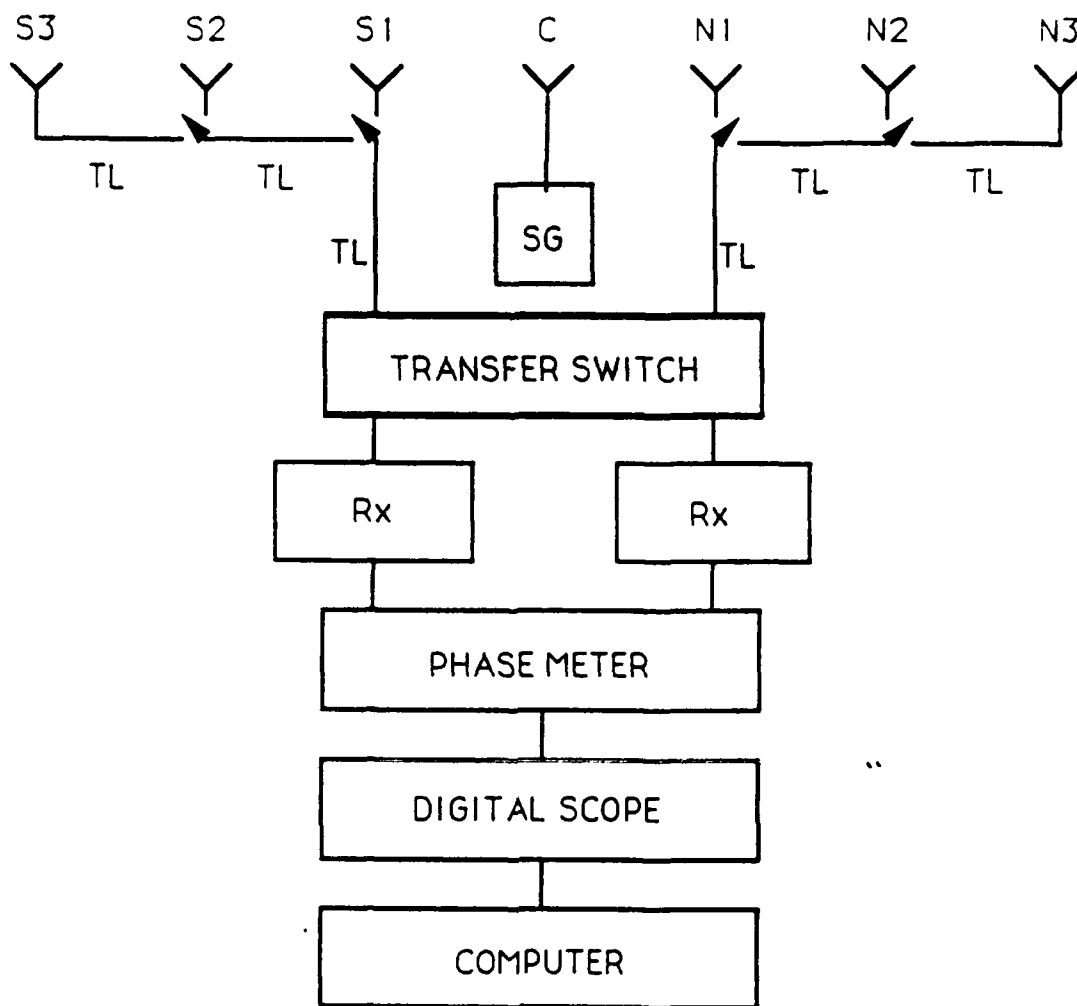
Rx: Two - Racal R-2174

Phase Meter: HP 3575A

Digital Scope: Tektronix 7D20

Computer: Mac Plus (45M Storage)

Figure 2.2a East-West Baseline System Configuration



TL: RG-214 (145 m. Section)

SG: Signal Generator (Calibration)

Rx: Two - Racal R-2174

Phase Meter: HP 3575A

Digital Scope: Tektronix 7D20

Computer: Mac Plus (45M Storage)

Figure 2.2b North-South Baseline System Configuration

necessary for the calibration procedure.

2.2 Calibration Procedure

A "central antenna" calibration technique was used to calibrate out any 2-element phase difference error components, eg. ground conductivity variations, transmission line electrical length deviations, etc. The local calibration signal power (milliwatts) is set to simulate the received signal power of ionospheric measurements as indicated on the receiver RF signal strength bargraph indicator. The basic concept is to transmit the calibration signal from a geometrically-centered element and measure the phase difference between two symmetrically spaced receive elements, eg. $\text{Phase(A)} - \text{Phase(B)}$. If the measurement is repeated with the two antenna/transmission line paths interchanged (via RF DPDT transfer switch), a second phase difference can be measured, eg. $\text{Phase(B)} - \text{Phase(A)}$. Hence, the mean of these two measurements represents the zero phase difference reference associated with cable length or ground conductivity differences in the actual ionospheric measurements. Figure 2.3 is a typical calibration plot. Since a commercially-available transfer switch had not been procured during these measurements, simple connection interchange was used which took approximately 20 seconds to perform. Hence, Figure 2.3 indicates a 20 second dwell in one configuration, a 20 second interchange period (meaningless data), another 20 second dwell in the interchanged configuration, another 20 interchange period (meaningless data), and, finally, another 20 second dwell in the original configuration. Hence, the phase difference zero reference is the mean of the two outer regions and the inner region. Since this is not automated into the measurement, this phase difference zero reference is determined manually from the data.

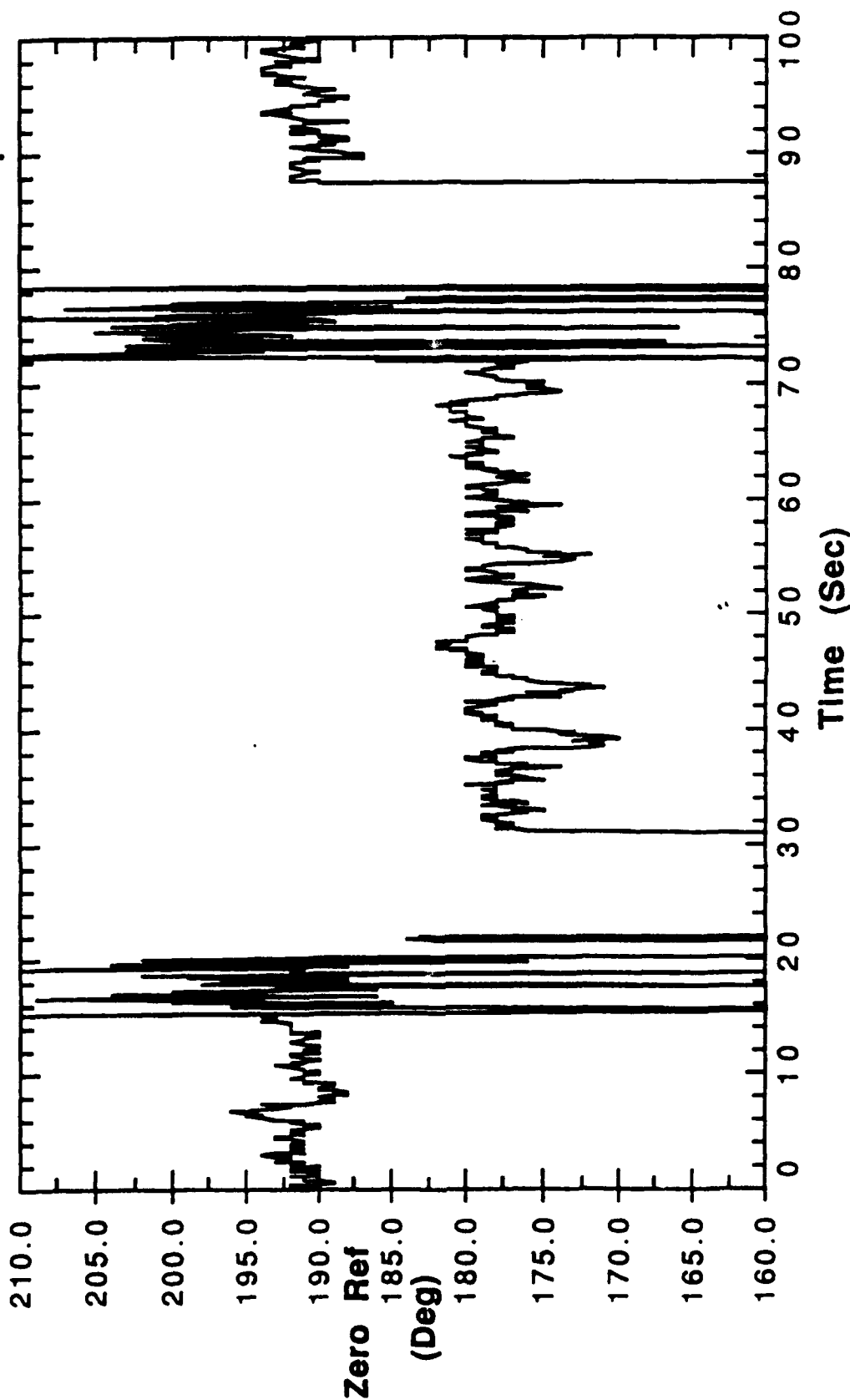
2.3 CW E-W Baseline Ionospheric Measurements

Initial Ava, NY-FAU measurements used unmodulated CW at 10.205 and 20.861 MHz with a 400 Hz receiver bandwidth. No mode separation techniques were included with these measurements. Since these were daytime measurements and path attenuation reduces

E-W FIELD SITE - 10FT VERTICALS

20.861 MHz, CW, E2/W2 SEPARATION = 580 m CALIBRATION

Date: 08-16-1990 Time: 22:21:56 Zero Phase: 0 Slope: 1



Resolution (Sec): .1

Data File: aug16.19

Figure 2.3

as frequency increases, the 20.861 MHz signal was much stronger than the 10.205 MHz and provided more meaningful results. Although the 10.205 MHz signal was detectable, the excessive amplitude fading did not allow any meaningful phase difference measurements which was an indication of multipath/multihop.

2.3.1 E1/W1 (Separation = 290 m.)

Figure 2.4 is an ionospheric measurement and Figure 2.5 is the associated calibration data. In Figure 2.5, the 0-15 second and 85-100 second periods represent Phase(W1) - Phase(E1) calibration signal and the 41-68 second period represents Phase(E1) - Phase(W1) calibration signal. Again, the phase difference zero reference is the mean of this data (57.5°). Some background noise is superimposed on these levels and would also be present on ionospheric data. It is noted the vertical scale differences between ionospheric data and calibration data. Figure 2.6 represents a histogram of the Figure 2.4 data. Figure 2.7 illustrates the geometry of the configuration, the angle-of-arrival ambiguities, and ambiguity resolution based on standard map measurements. Again, the ambiguity number n is the integral number of electrical wavelengths included in the geometric length $L1$.

2.3.2 E2/W2 (Separation = 580 m.)

Figures 2.8-2.10 represent similar plots as those described above except for E2/W2 measurements. Figure 2.11 depicts the geometry, angle-of-arrival ambiguities, and comparison with known direction-of-arrival.

2.3.3 W1/E2 (Separation = 435 m.)

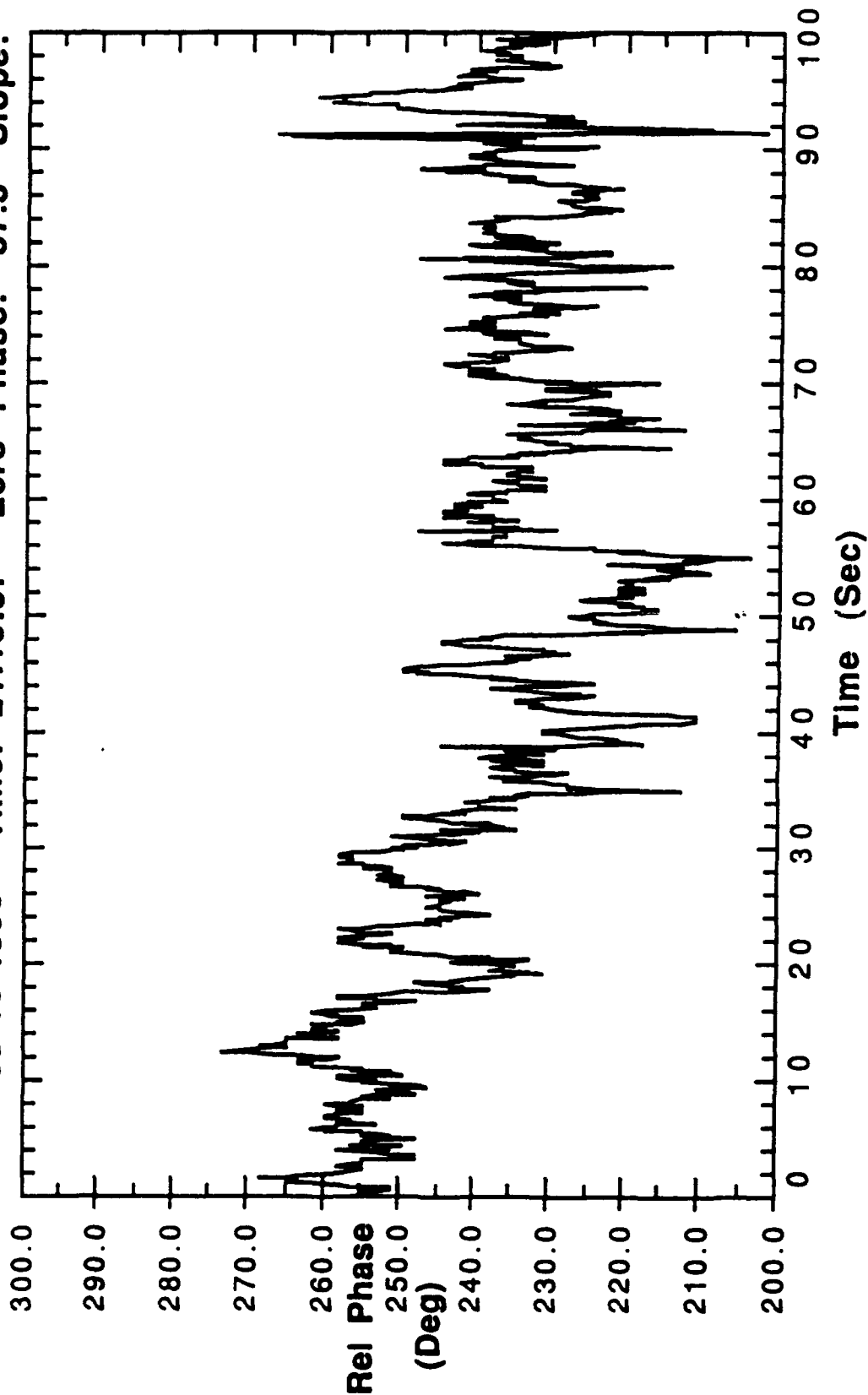
Due to lack of symmetry of two elements relative to the central calibration element, the "central antenna" calibration technique cannot be used for obtaining a phase difference zero reference and, hence, mean angle-of-arrival. Nevertheless, the time-varying component of phase difference is valid and is shown in Figures 2.12 and 2.13.

2.3.4 W2/E1 (Separation = 435 m.)

E-W FIELD SITE - 10FT VERTICALS

20.861 MHz, CW, E1/W1 SEPARATION = 290 m DATA RUN

Date: 08-16-1990 Time: 21:15:37 Zero Phase: 57.5 Slope: 2.1157



Resolution (Sec): .1

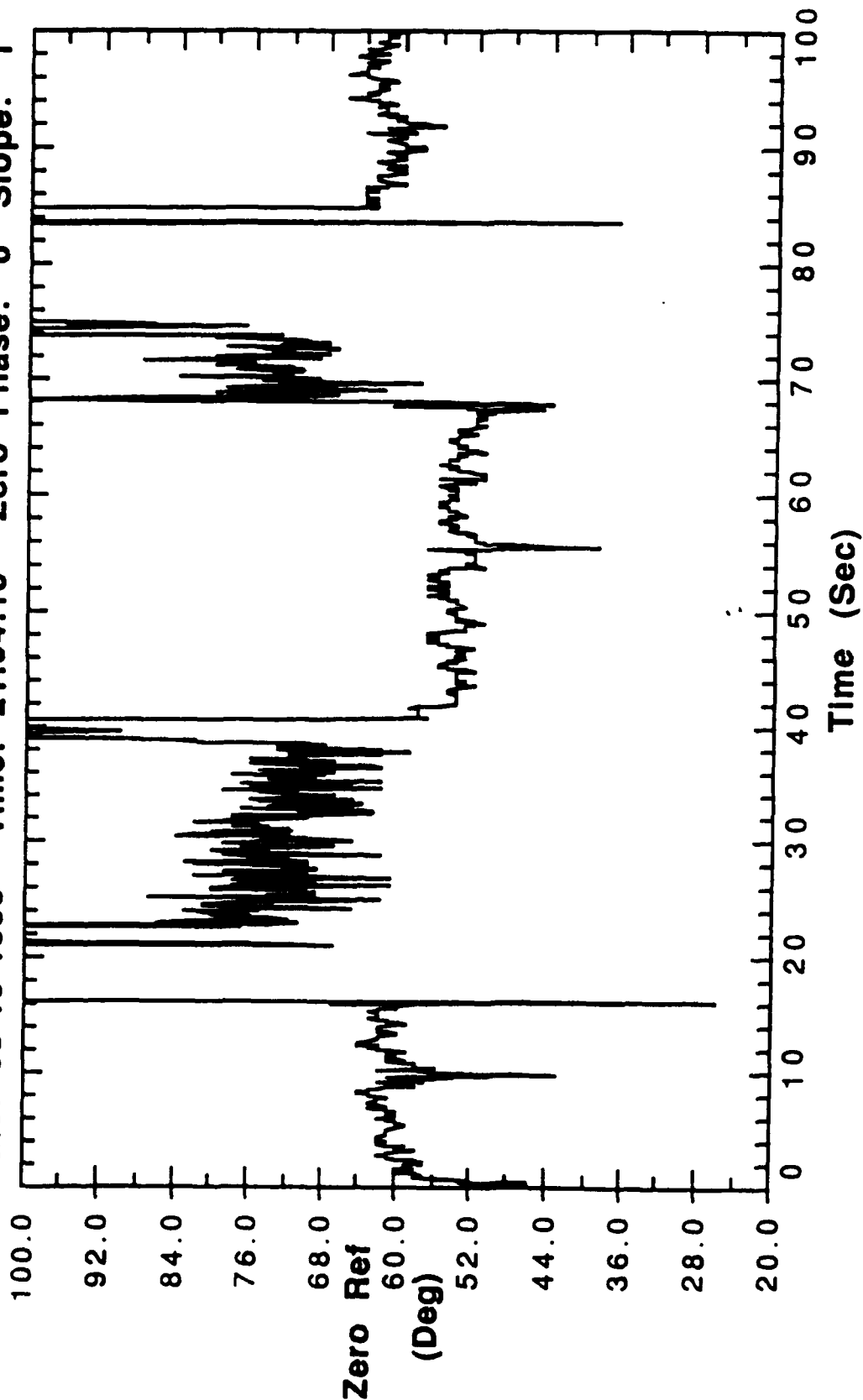
Data File: aug16.11

Figure 2.4

E-W FIELD SITE - 10FT VERTICALS

20.861 MHz, CW, E1/W1 SEPARATION = 290 m CALIBRATION

Date: 08-16-1990 Time: 21:04:10 Zero Phase: 0 Slope: 1



Resolution (Sec): .1

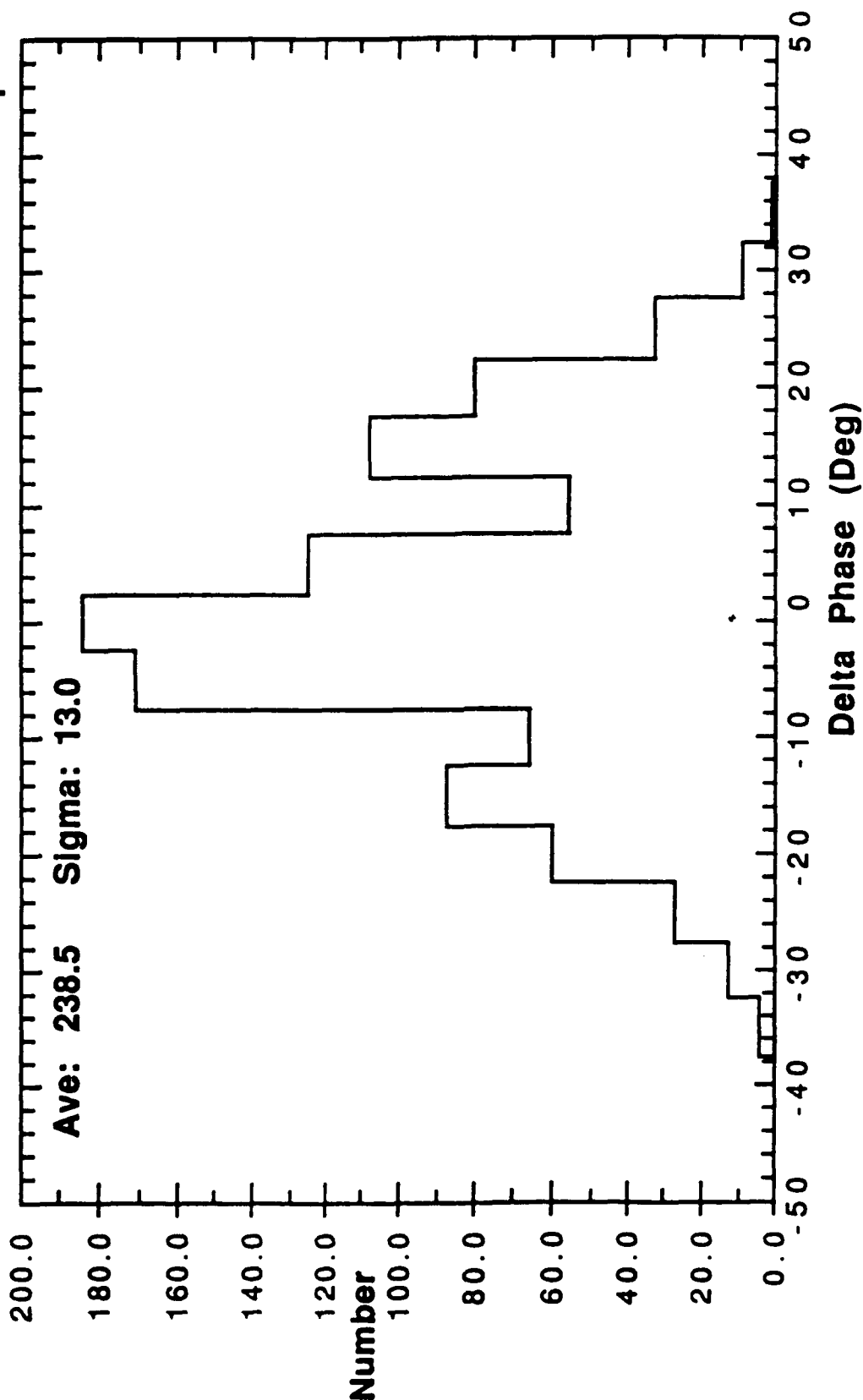
Data File: aug16.09

Figure 2.5

E-W FIELD SITE - 10FT VERTICALS

20.861 MHz, CW, E1/W1 SEPARATION = 290 m DATA RUN

Date: 08-16-1990 Time: 21:15:37 Zero Phase: 57.5 Slope: 2.1157

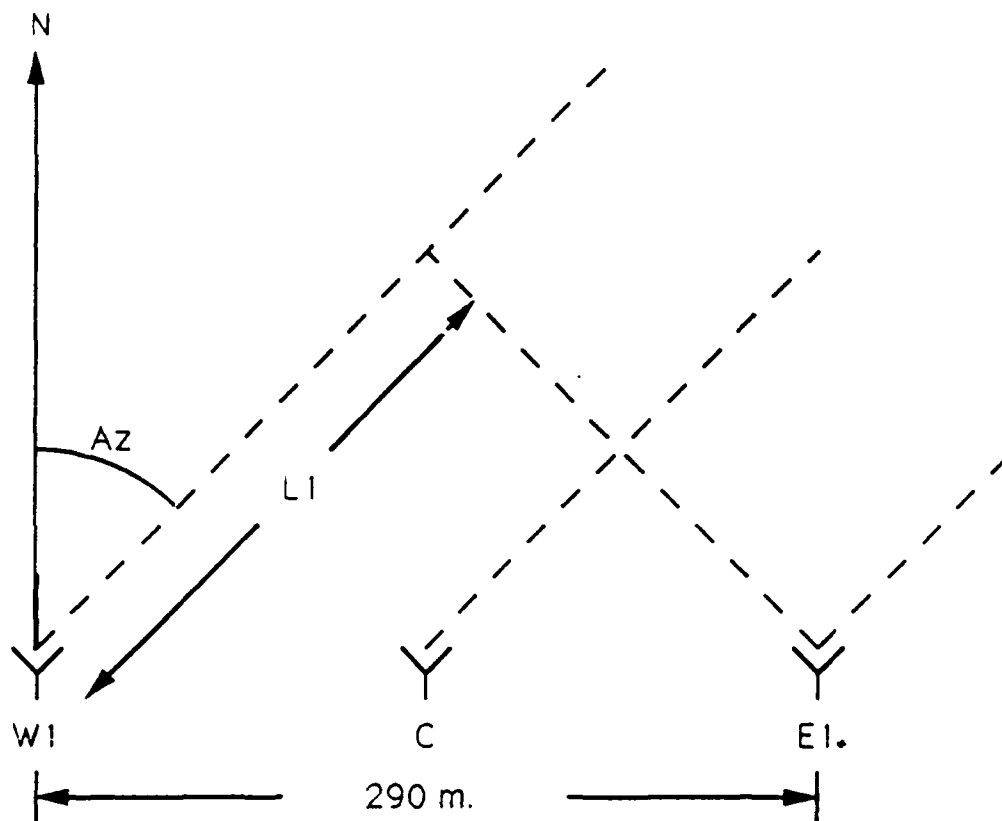


Resolution (Sec): .1

Bin Width (Deg): 5.0

Data File: aug16.11

Figure 2.6



$$\lambda = c / (20.861 \text{ MHz}) = 14.38 \text{ m.}$$

$$\text{Mean}[\text{Phase}(W1) - \text{Phase}(E1)] = +238.5^\circ = -121.5^\circ \text{ (measured)}$$

$$\text{Phase}(W1) - \text{Phase}(E1) = -\beta(L1)$$

$$-n(2\pi) - 121.5^\circ (\pi / 180^\circ) = -(2\pi / 14.38 \text{ m.}) L1$$

$$L1 = 4.85 \text{ m.} + n(14.38 \text{ m.})$$

$$\text{Az} = \arcsin(L1 / 290 \text{ m.})$$

n	0	1	2	3	4	5
Az(°)	1.0	3.8	6.7	9.5	12.4	15.3

Figure 2.7a Angle-of-Arrival Azimuth Calculations

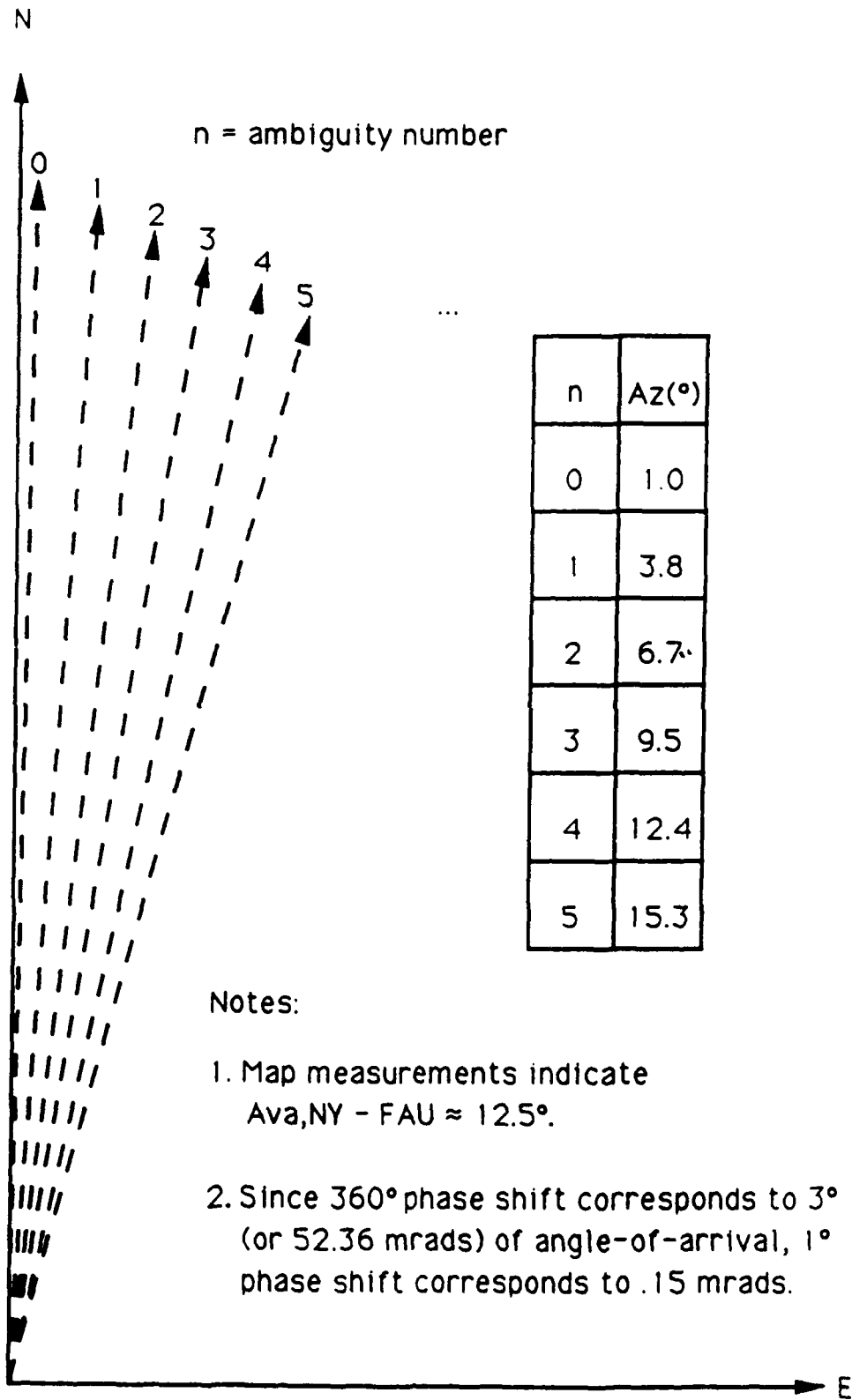
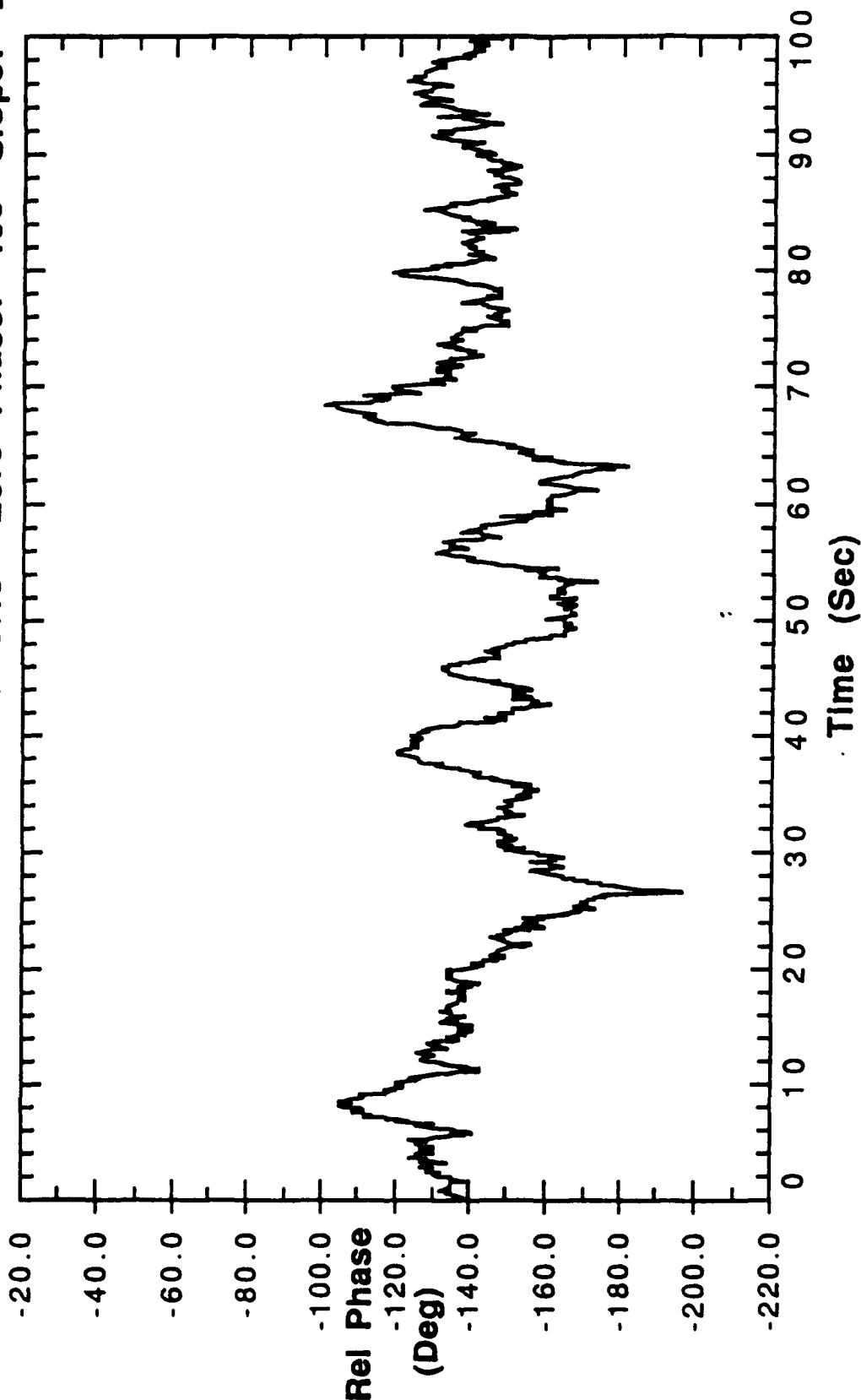


Figure 2.7b Angle-of-Arrival Azimuth Ambiguities

E-W FIELD SITE - 10FT VERTICALS

20.061 MHz, CW, E2/W2 SEPARATION = 580 m DATA RUN

Date: 08-16-1990 Time: 20:16:19 Zero Phase: 195 Slope: 2.1157



Resolution (Sec): .1

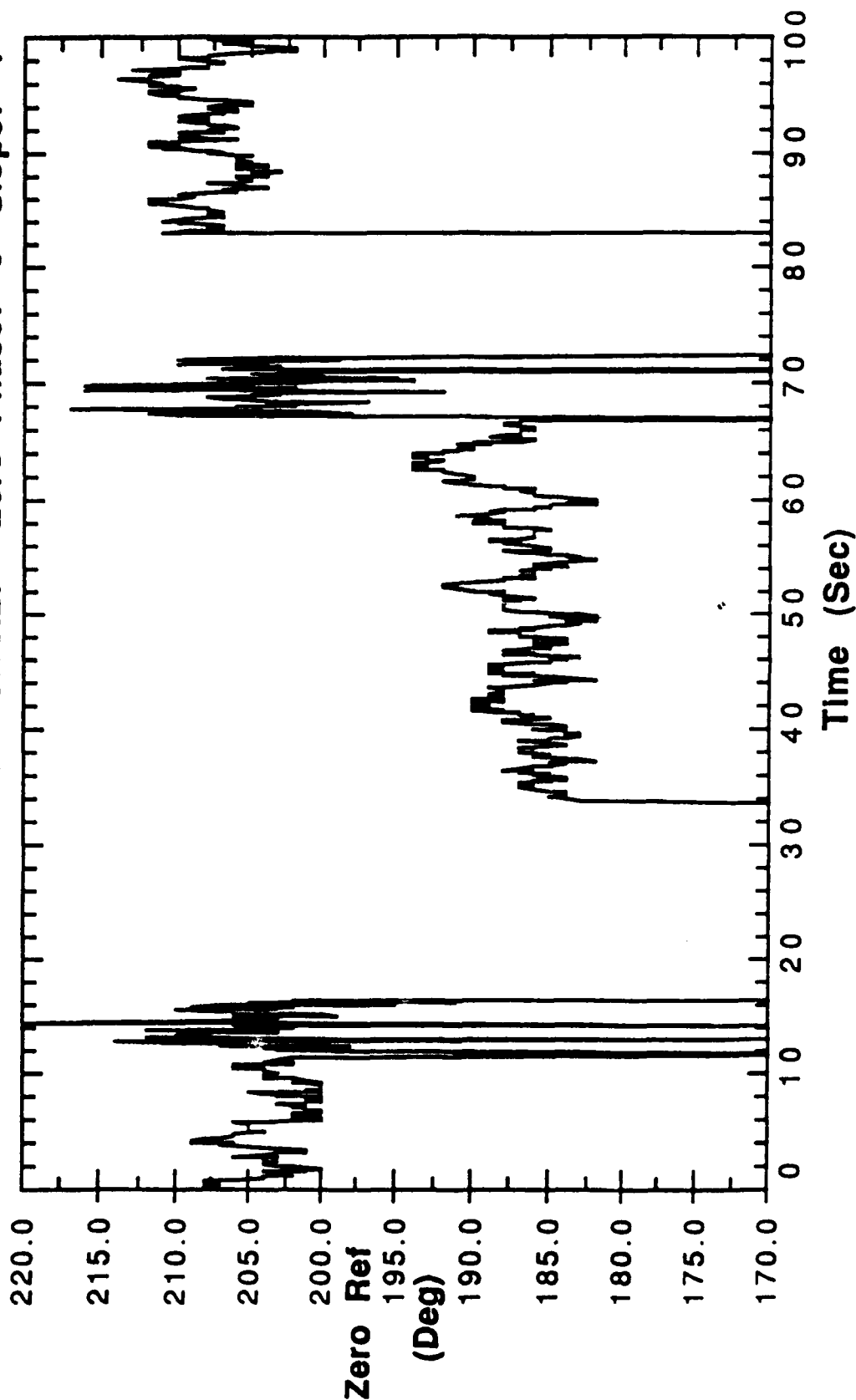
Data File: aug16.04

Figure 2.8

E-W FIELD SITE - 10FT VERTICALS

20.061 MHz, CW, E2/W2 SEPARATION = 580 m CALIBRATION

Date: 08-16-1990 Time: 20:08:27 Zero Phase: 0 Slope: 1



Resolution (Sec): .1

Data File: aug16.03

Figure 2.9

E-W FIELD SITE - 10FT VERTICALS

20.061 MHz, CW, E2/W2 SEPARATION = 580 m DATA RUN

Date: 08-16-1990 Time: 20:16:19 Zero Phase: 195 Slope: 2.1157

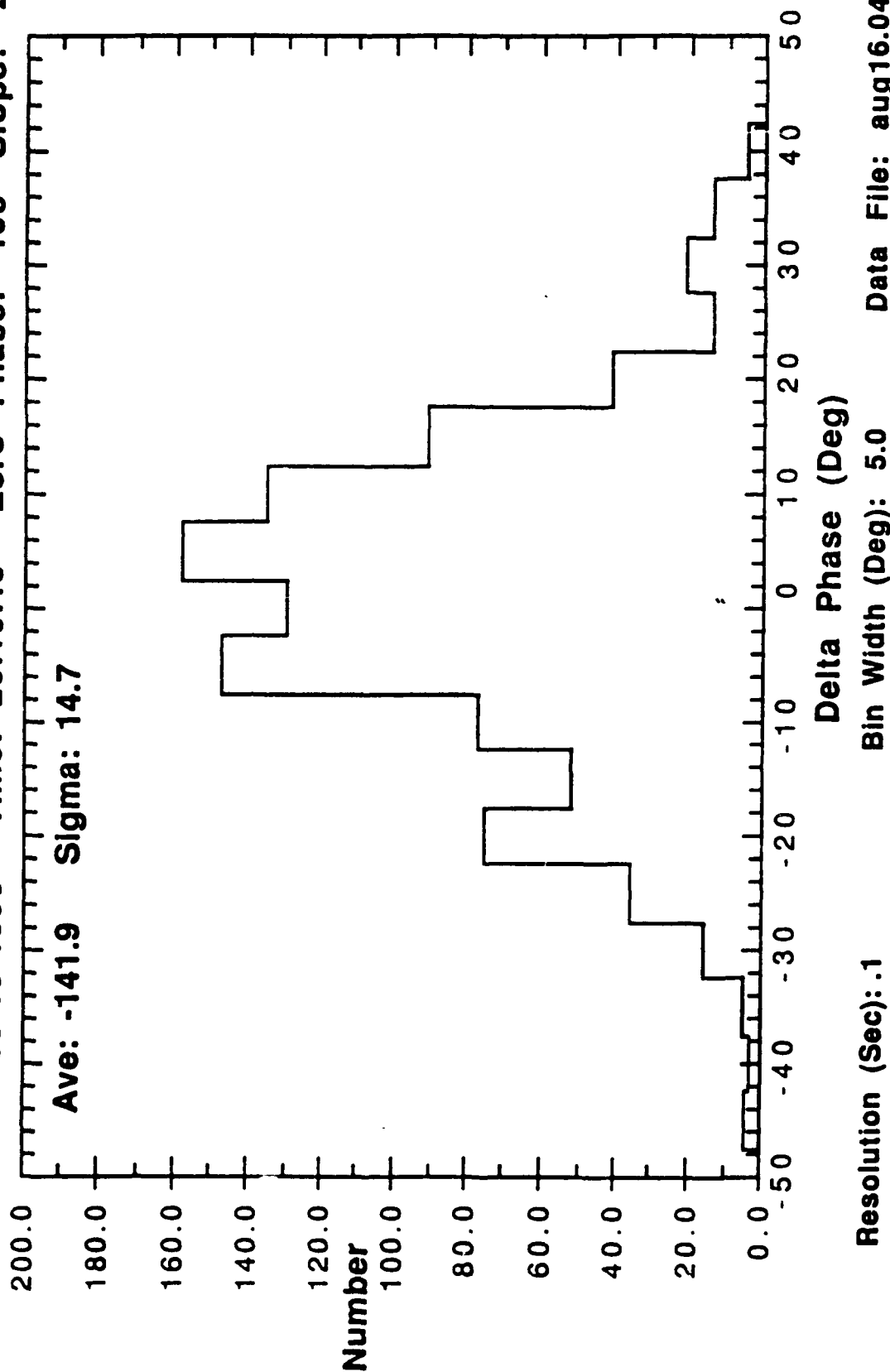
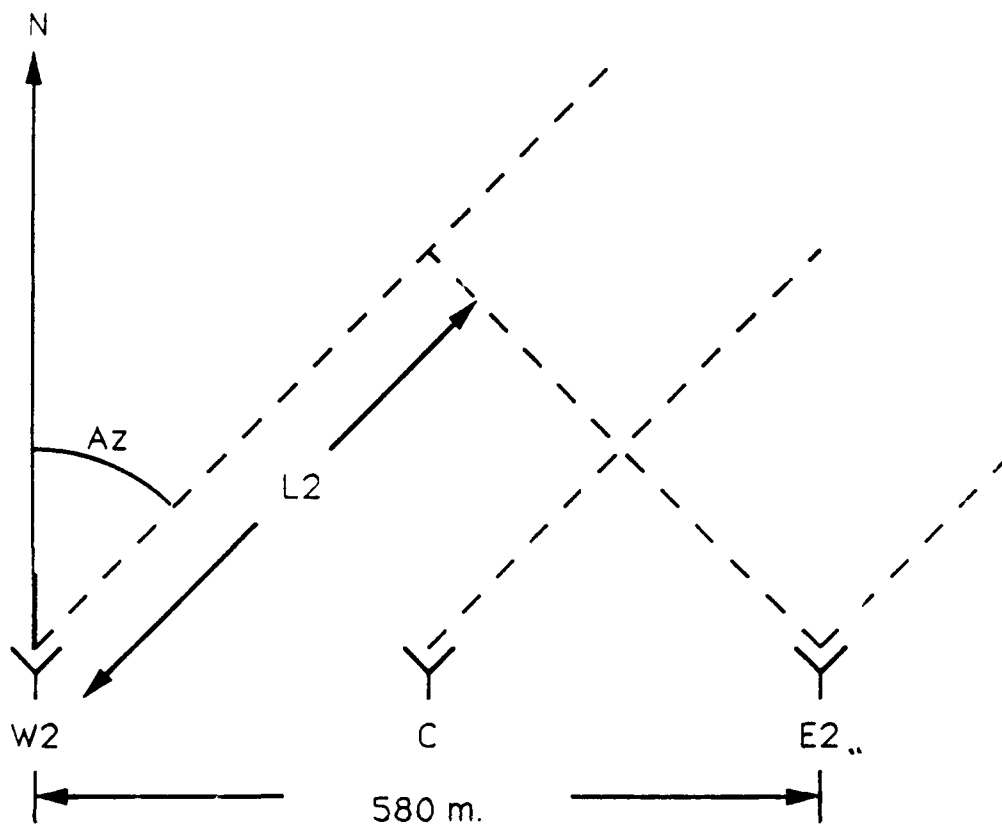


Figure 2.10



$$\lambda = c / (20.861 \text{ MHz}) = 14.38 \text{ m.}$$

$$\text{Mean}[\text{Phase}(W2) - \text{Phase}(E2)] = -142^\circ \text{ (measured)}$$

$$\text{Phase}(W2) - \text{Phase}(E2) = -\beta(L2)$$

$$-n(2\pi) - 142^\circ (\pi / 180^\circ) = -(2\pi / 14.38 \text{ m.}) L2$$

$$L2 = 5.61 \text{ m.} + n(14.38 \text{ m.})$$

$$Az = \arcsin(L2 / 580 \text{ m.})$$

n	0	1	2	3	4	5	6	7	8	9
Az(°)	0.6	2.0	3.4	4.8	6.2	7.7	9.1	10.6	12.0	13.5

Figure 2.11a Angle-of-Arrival Azimuth Calculations

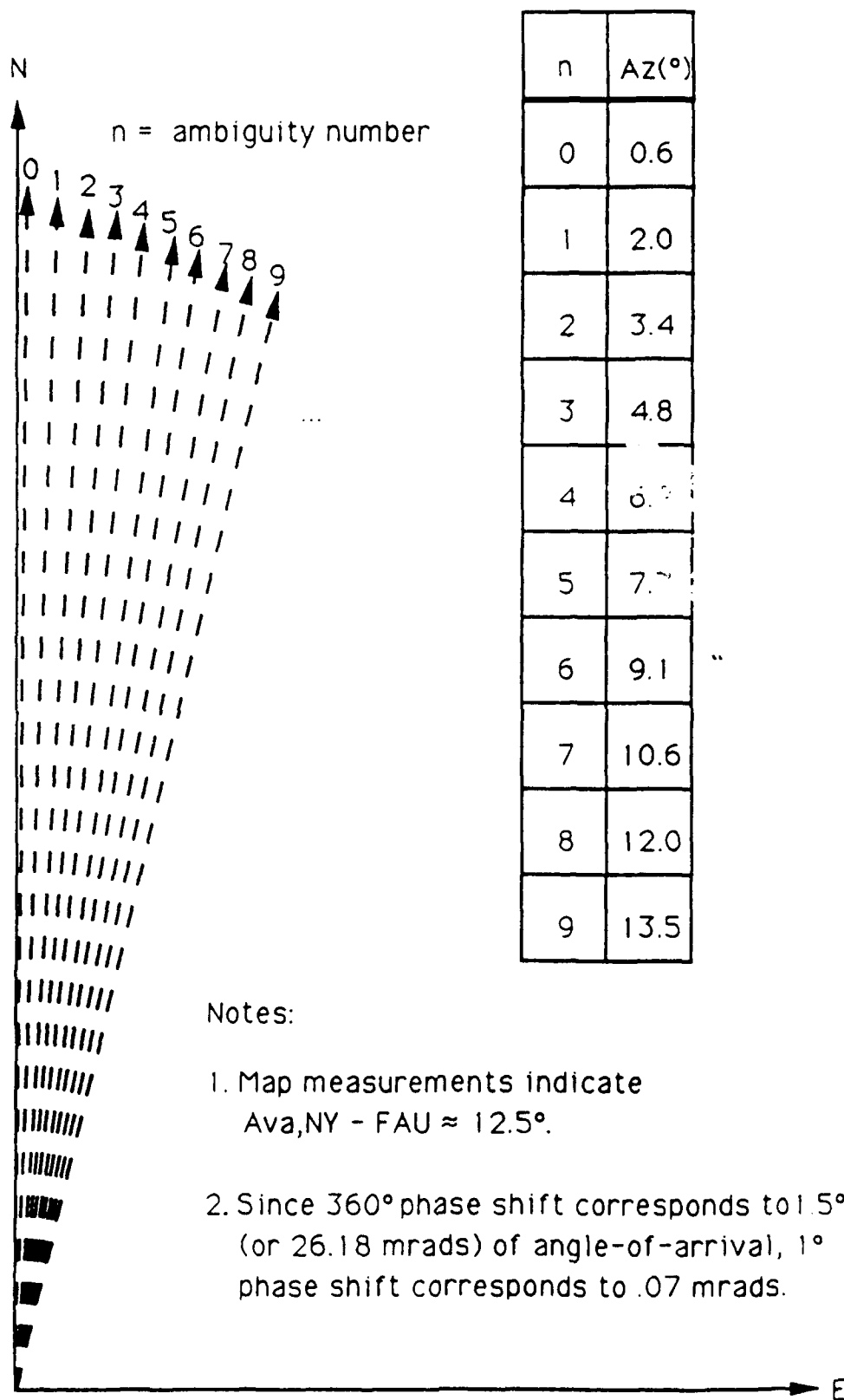
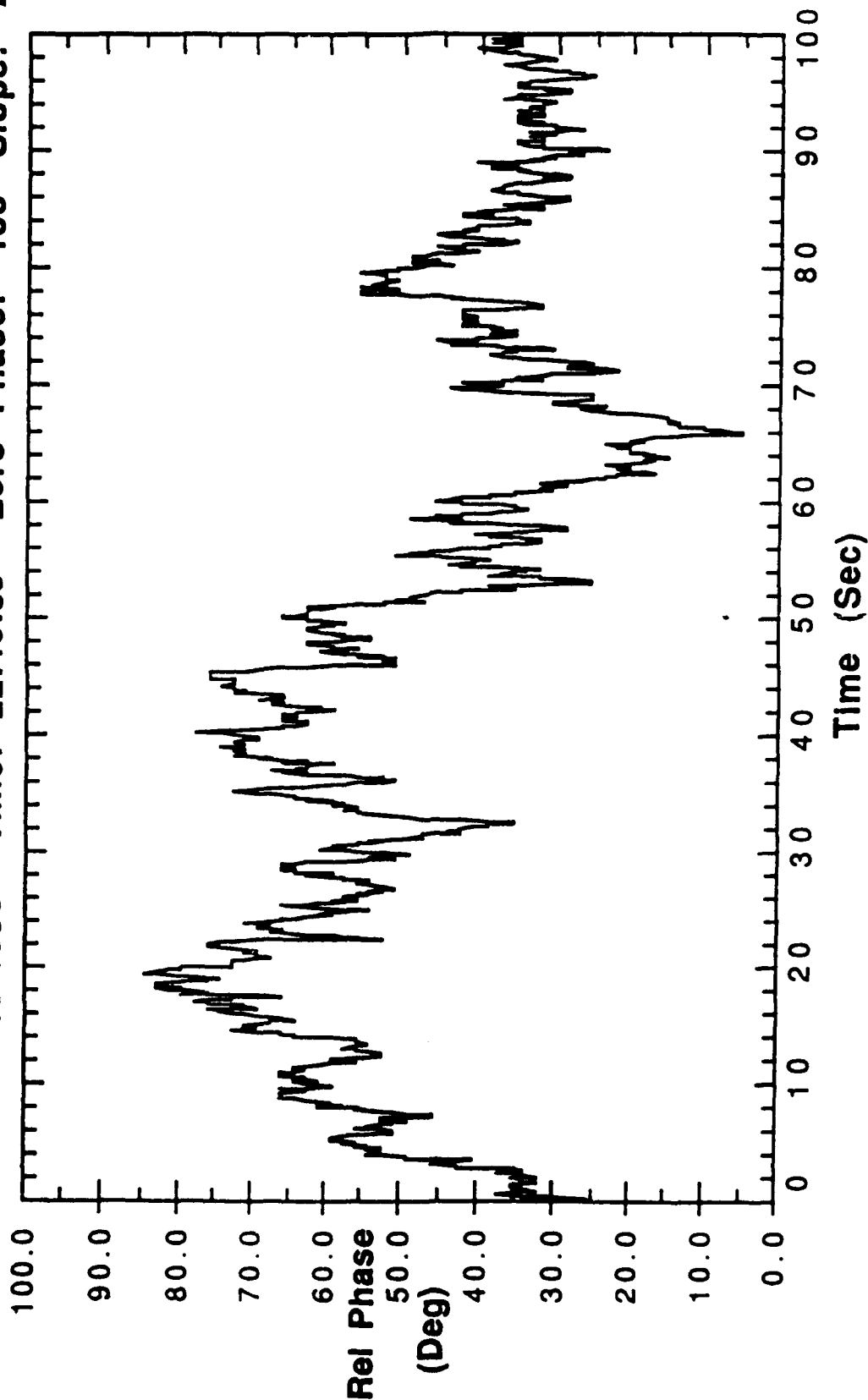


Figure 2.11b Angle-of-Arrival Azimuth Ambiguities

E-W FIELD SITE - 10FT VERTICALS

20.861 MHz, CW, E2/W1 SEPARATION = 435 m DATA RUN

Date: 08-16-1990 Time: 22:46:39 Zero Phase: 185 Slope: 2.1157



Resolution (Sec): .1

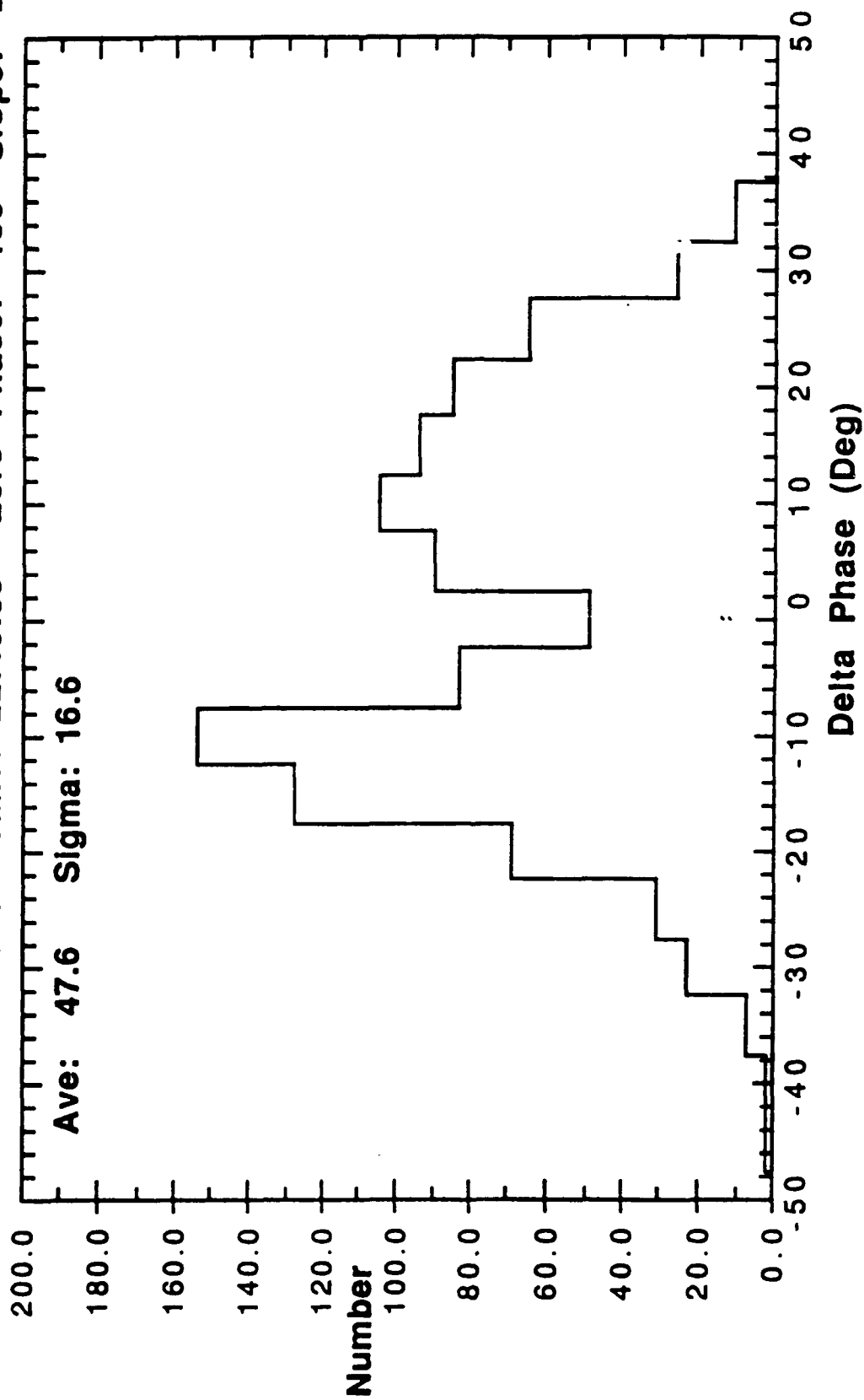
Data File: aug16.21

Figure 2.12

E-W FIELD SITE - 10FT VERTICALS

20.861 MHz, CW, E2/W1 SEPARATION = 435 m DATA RUN

Date: 08-16-1990 Time: 22:46:39 Zero Phase: 185 Slope: 2.1157



Resolution (Sec): .1 Bln Width (Deg): 5.0 Data File: aug16.21

Figure 2.13

Again, the "central antenna" calibration technique cannot be used and only the time-varying component of phase difference is valid and shown in Figures 2.14 and 2.15.

2.3.5 Summary

(Representative Data With Amplitude Within Receiver Dynamic Range)

Configuration	StdDev [Phase Difference] (degrees)	StdDev [Azimuth] (mrads)	Mean [Azimuth] (degrees)
W1/E1 (290 m.)	13.0	1.95	12.4
W1/E2 (435 m.)	16.6	1.87	-
W2/E1 (435 m.)	41.4	4.66	-
W2/E2 (580 m.)	14.7	1.10	12.0
W2/E2 (580 m.)	22.1	1.65	12.1
W2/E2 (580 m.)	50.6	3.79	12.6

Although the angle-of-arrival data is reasonable when compared to the known direction-of-arrival, standard deviation of phase difference does not indicate a trend as aperture width increases. This could be due to an insufficient number of samples or the lack of any mode separation techniques in the unmodulated CW transmissions. Hence, further measurements must incorporate these mode separation techniques.

2.4 N-S Baseline Ionospheric Measurements

Again, since the 20.861 MHz signals were much stronger than the 10.205 MHz signals, only 20.861 MHz results are reported.

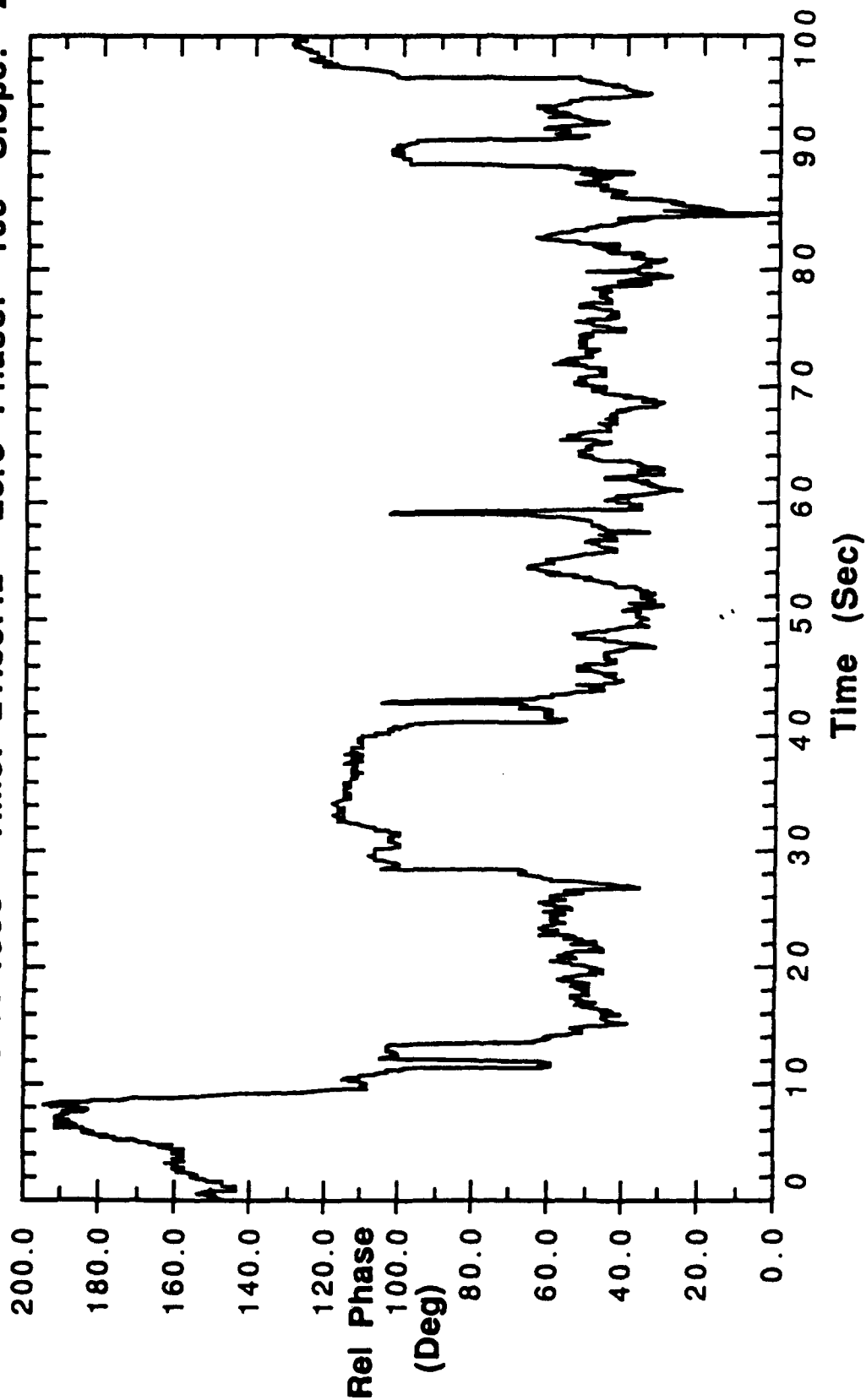
2.4.1 S1/N1 (Separation = 290 m.)

Figures 2.16-2.18 are these measurement results and Figure 2.19 depicts the associated geometry and calculations.

E-W FIELD SITE - 10FT VERTICALS

20.861 MHz, CW, E1/W2 SEPARATION = 435 m DATA RUN

Date: 08-16-1990 Time: 21:55:42 Zero Phase: 109 Slope: 2.1157



Resolution (Sec): .1

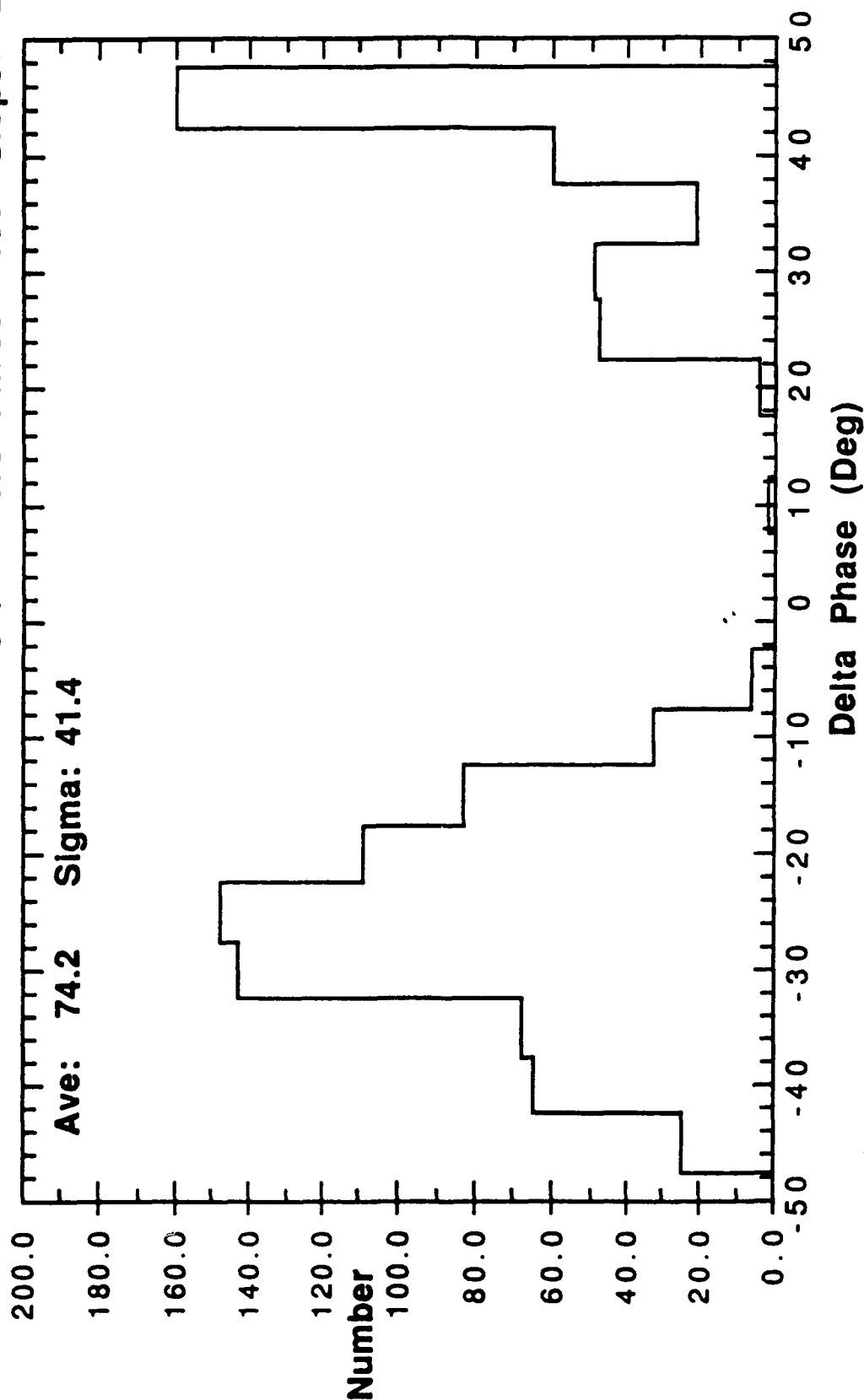
Data File: aug16.16

Figure 2.14

E-W FIELD SITE - 10FT VERTICALS

20.861 MHz, CW, E1/W2 SEPARATION = 435 m DATA RUN

Date: 08-16-1990 Time: 21:55:42 Zero Phase: 109 Slope: 2.1157



Resolution (Sec): .1

Bin Width (Deg): 5.0

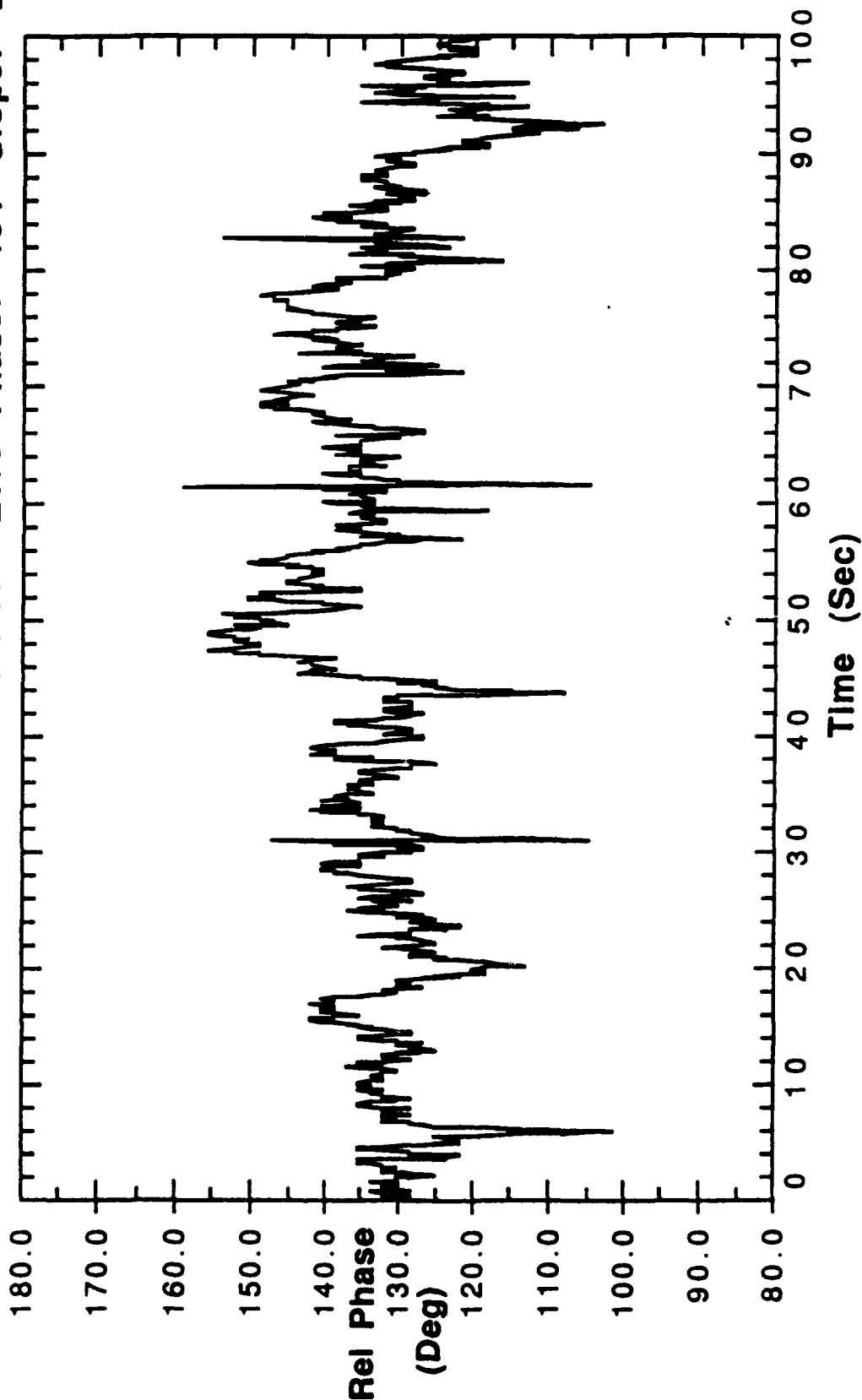
Data File: aug16.16

Figure 2.15

N-S FIELD SITE-10 FT VERTICALS

20.061 MHz, CW, N1/S1 SEPARATION = 290 m DATA RUN

Date: 08-24-1990 Time: 20:45:57 Zero Phase: 104 Slope: 2.1157



Resolution (Sec): .1

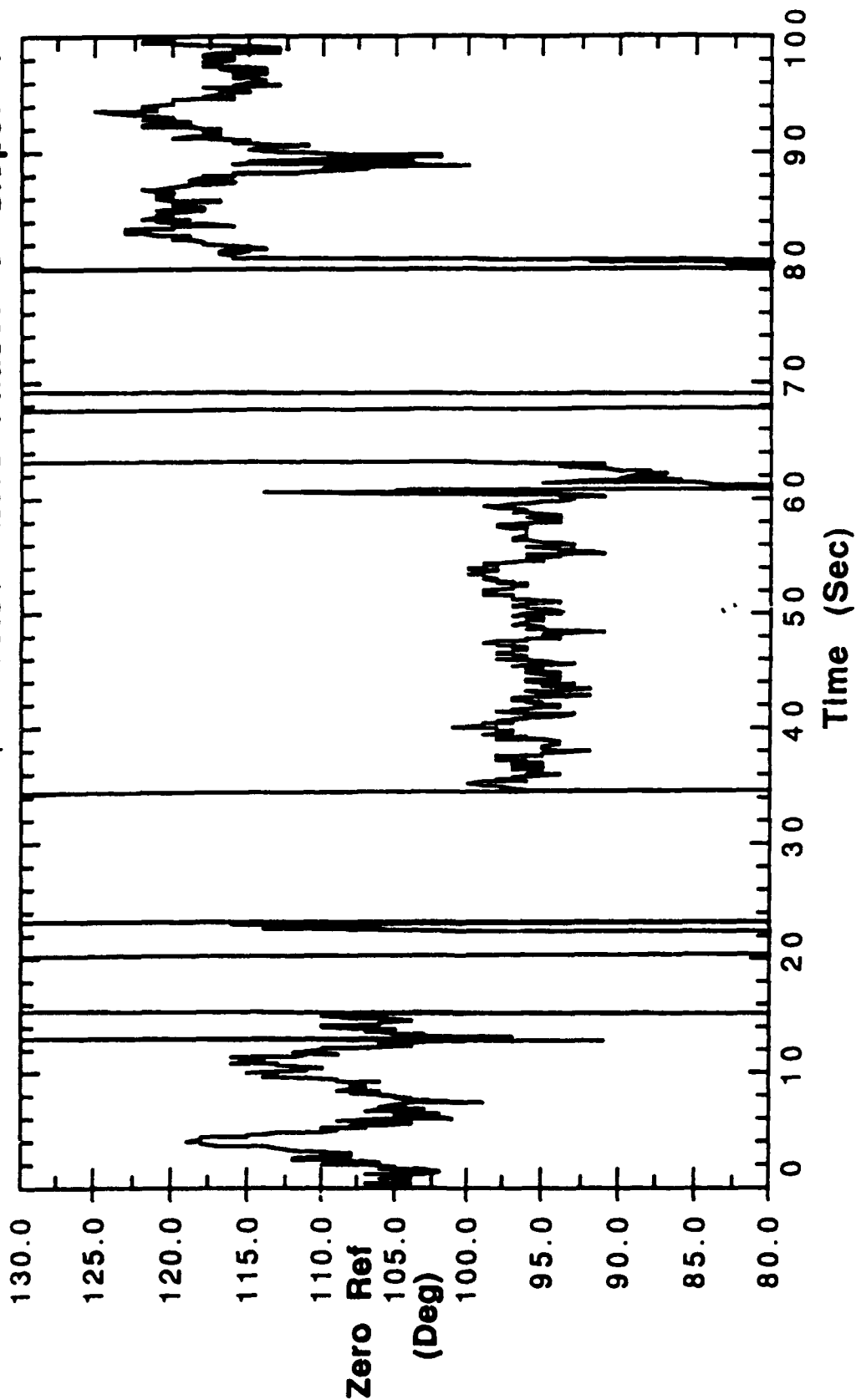
Data File: aug24.04

Figure 2.16

N-S FIELD SITE-10 FT VERTICALS

20.061 MHz, CW, N1/S1 SEPARATION = 290 m CALIBRATION

Date: 08-24-1990 Time: 20:36:31 Zero Phase: 0 Slope: 1



Resolution (Sec): .1

Data File: aug24.03

Figure 2.17

N-S FIELD SITE-10 FT VERTICALS

20.061 MHz, CW, N1/S1 SEPARATION = 290 m DATA RUN

Date: 08-24-1990 Time: 20:45:57 Zero Phase: 104 Slope: 2.1157

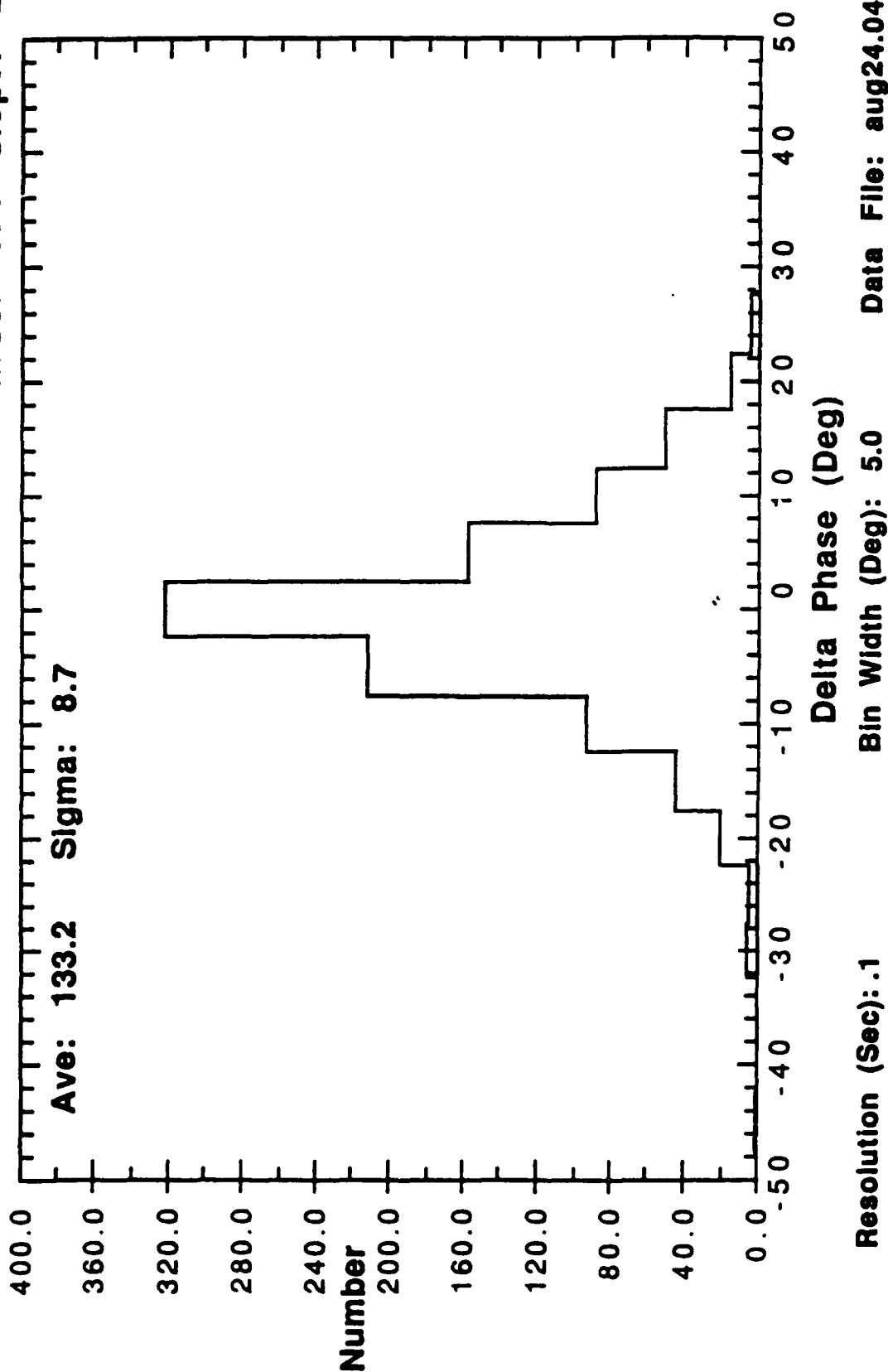
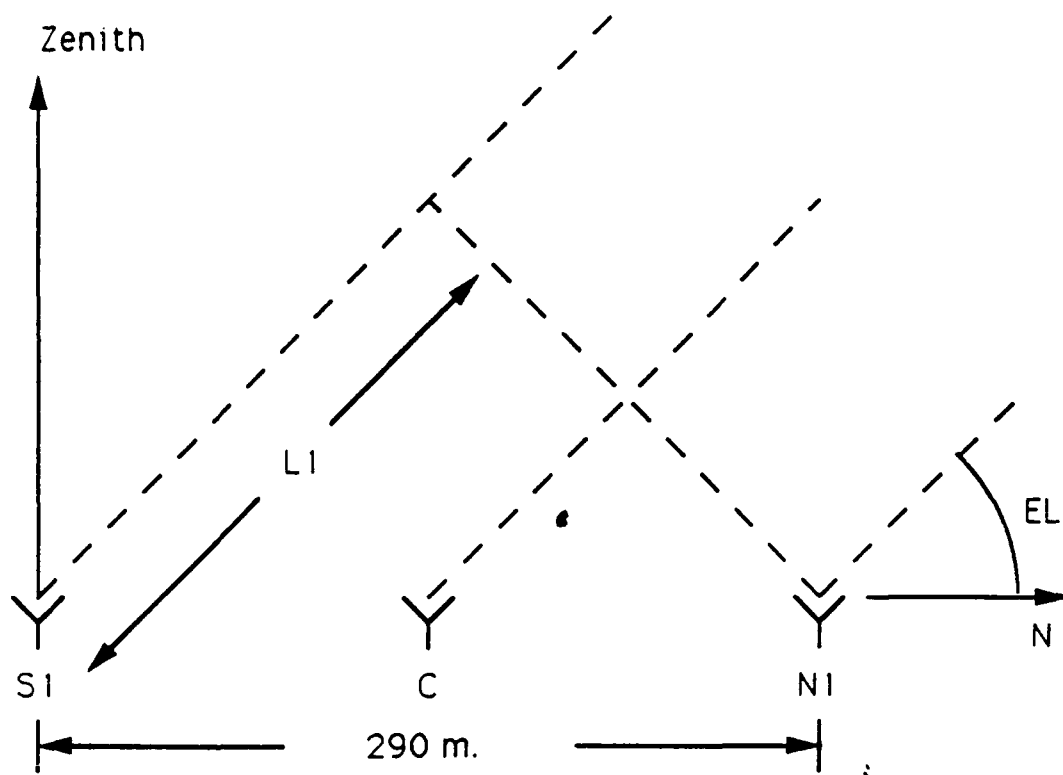


Figure 2.18



$$\lambda = c / (20.861 \text{ MHz}) = 14.38\text{m.}$$

$$\text{Mean}[\text{Phase}(S1) - \text{Phase}(N1)] = -227^\circ \text{ (measured)}$$

$$\text{Phase}(S1) - \text{Phase}(N1) = -\beta(L1)$$

$$-n(2\pi) - 227^\circ (\pi / 180^\circ) = -(2\pi / 14.38\text{m.})L1$$

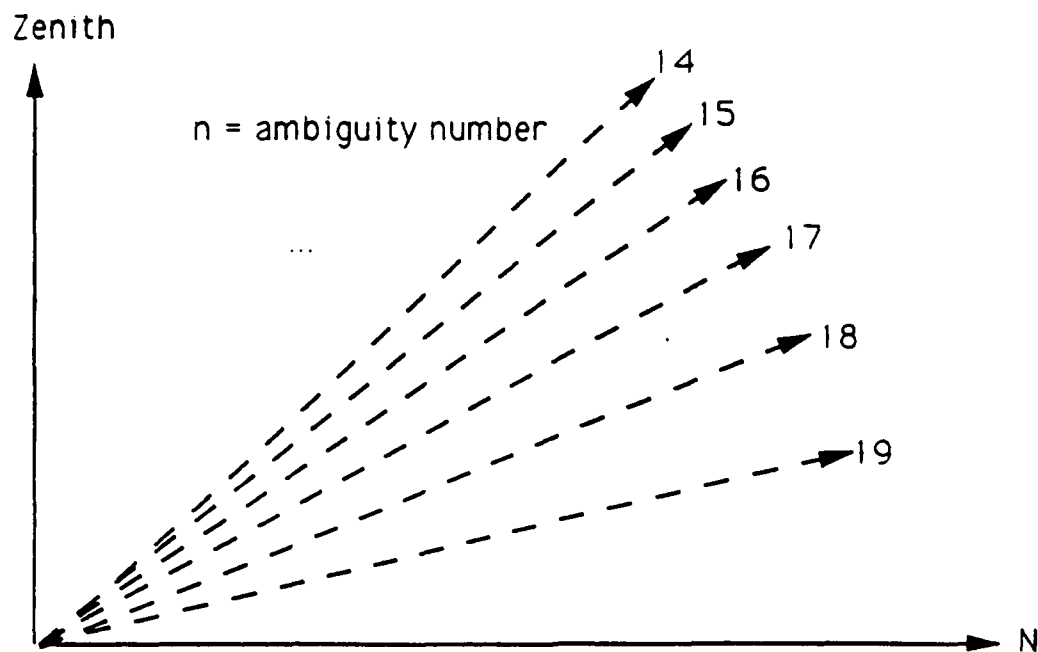
$$L1 = 9.06\text{m.} + n(14.38\text{m.})$$

$$EL = \arccos(L1 / 290\text{m.})$$

Note: Since $L1 \leq 290\text{m.}$, $n \leq 19$.

n	19	18	17	16	15	14
EL(°)	13.3	22.5	29.0	34.5	39.2	43.5

Figure 2.19a Angle-of-Arrival Elevation Calculations



n	19	18	17	16	15	14
EL(°)	13.3	22.5	29.0	34.5	39.2	43.5

Notes:

1. For the Ava,NY-FAU ground range of 2400km., Griffiths [33] indicates elevation angle-of-arrival ranges 6° – 12° for effective heights of 200–400km., respectively.
2. Since 360° phase shift corresponds to 9.2° (or 160.5 mrad) of angle-of-arrival, 1° phase shift corresponds to .45 mrad.

Figure 2.19b Angle-of-Arrival Elevation Ambiguities

2.4.2 S2/N2 (Separation = 580 m.)

Figures 2.20-2.22 are these measurement results and Figure 2.23 depicts the associated geometry and calculations

2.4.3 Summary

(Representative Data With Amplitude Within Receiver Dynamic Range)

Configuration	Std Dev [Phase Difference] (degrees)	Std Dev [Elevation] (mrads)	Mean [Elevation] (degrees)
S1/N1 (290 m.)	8.7	4.6	13.3
S2/N2 (580 m.)	26.7	8.0	14.5

The standard deviation and mean of elevation follow expected results. When aperture was doubled, standard deviation of phase difference more than doubled and standard deviation of elevation degraded from constancy. The mean of elevation results agree closely with published models [34], i.e. 6° - 12° for virtual heights of 200-400 km.

3.0 Mode Separation Experiments

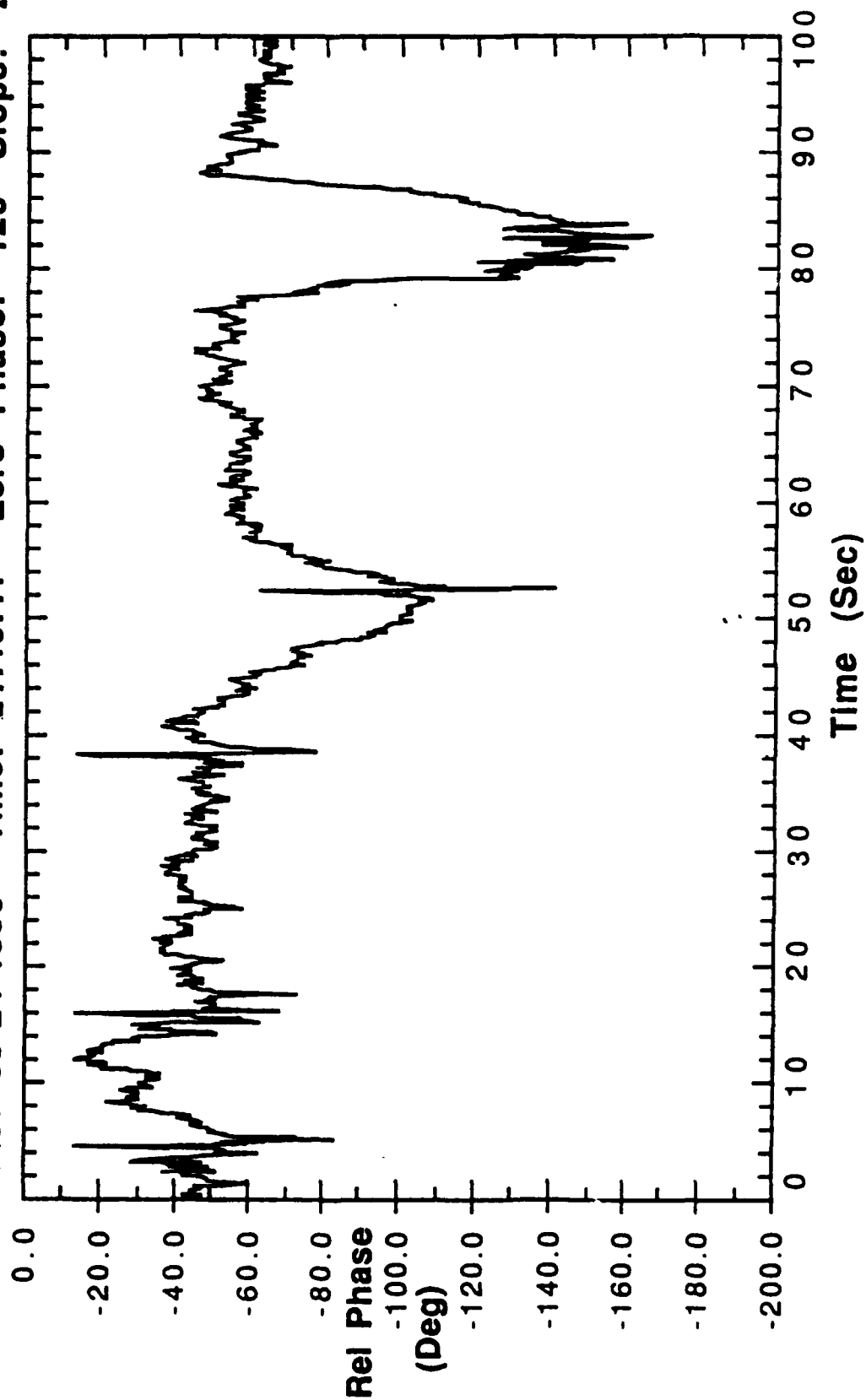
An analysis of the CW data indicate the presence of sudden phase jumps. In order to study the origin of these phase jumps, a novel mode separation technique [33], was employed.

The transmitter signal was phase modulated with a 127 bit length Pseudo Random (PN) code sequence (i.e. length 7 shift register with 1,7 feedback connections). The chip rate of the transmitted code was 10 KHz. At the receive site, two receivers were phase locked to a common 10 MHz reference signal. Normally, one would mix the received signal down to baseband and then correlate the baseband signal with the transmitted code sequence. However a technique based on direct IF sampling was used. The IF is directly undersampled and the conversion to baseband is done via software. This method eliminates the need for two balanced A/D converters needed in the normal process of conversion to baseband.

IN-3 FIELD DISE-IU FI VERTICALS

20.861 MHz, CW, N2/S2 SEPARATION = 580 m DATA RUN

Date: 08-24-1990 Time: 21:46:41 Zero Phase: 120 Slope: 2.1157



Resolution (Sec): .1

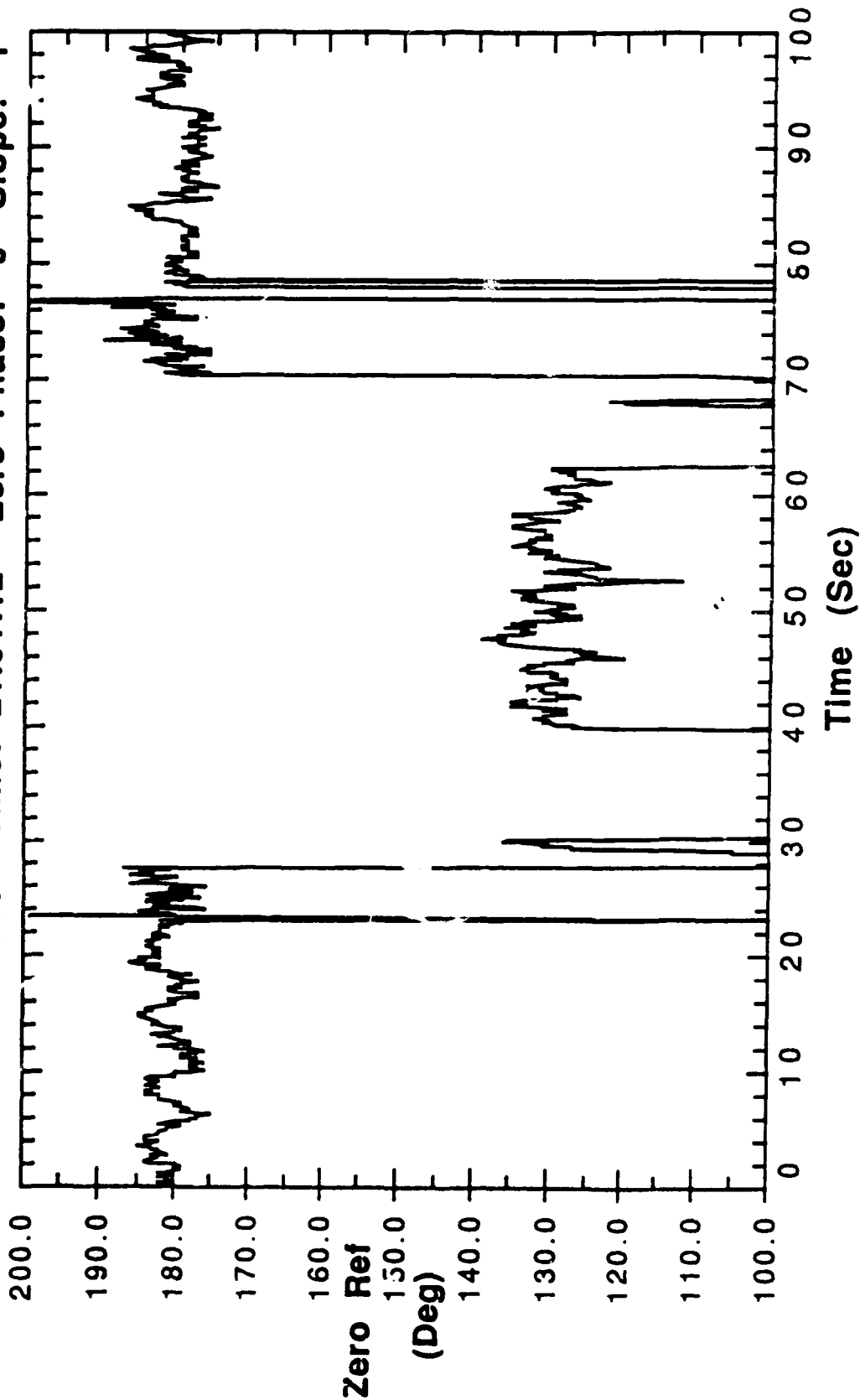
Data File: aug24.09

Figure 2.20

N-S FIELD SITE-10 FT VERTICALS

20.861 MHz, CW, N2/S2 SEPARATION = 580 m CALIBRATION

Date: 08-24-1990 Time: 21:51:12 Zero Phase: 0 Slope: 1



Resolution (Sec): .1

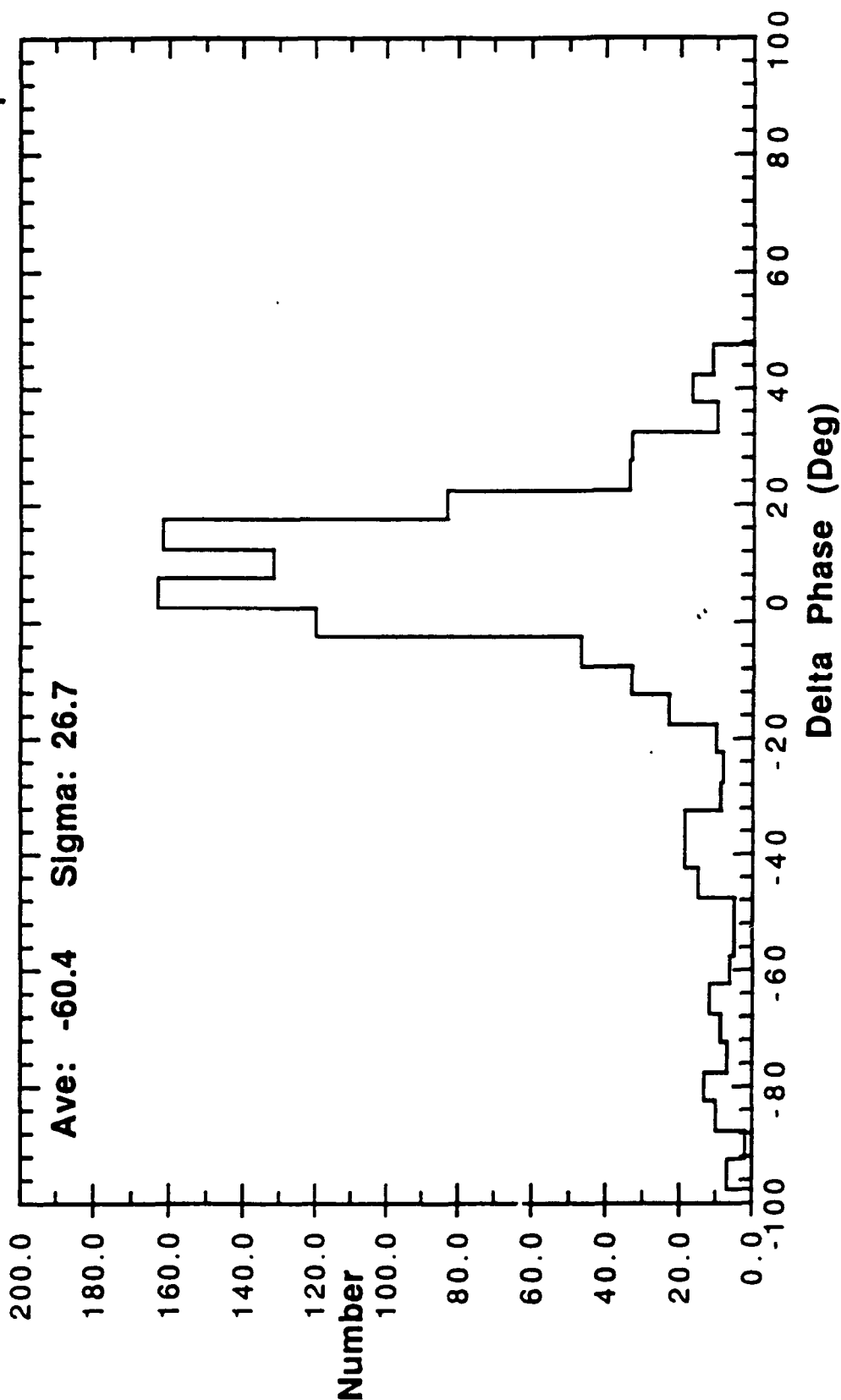
Data File: aug24.10

Figure 2.21

N-S FIELD SITE-10 FT VERTICALS

20.861 MHz, CW, N2/S2 SEPARATION = 580 m DATA RUN

Date: 08-24-1990 Time: 21:46:41 Zero Phase: 120 Slope: 2.1157

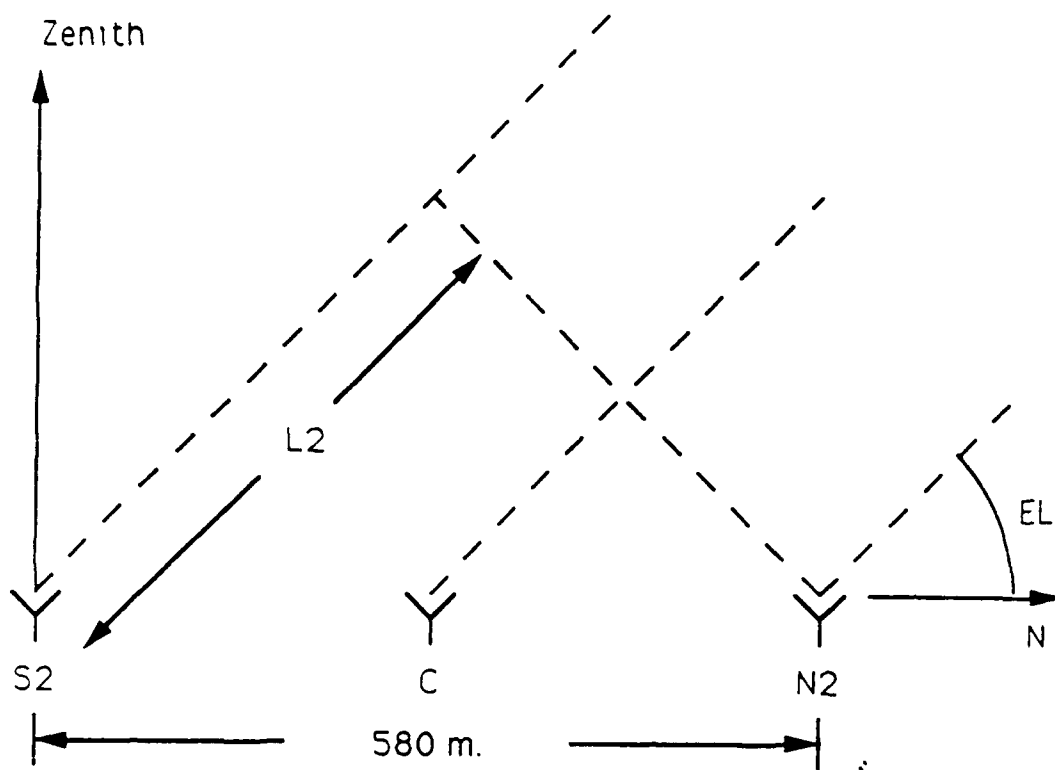


Resolution (Sec):.1

Bin Width (Deg): 5.0

Data File: aug24.09

Figure 2.22



$$\lambda = c / (20.861 \text{ MHz}) = 14.38 \text{ m.}$$

$$\text{Mean}[\text{Phase}(S2) - \text{Phase}(N2)] = -60.4^\circ \text{ (measured)}$$

$$\text{Phase}(S2) - \text{Phase}(N2) = -\beta(L2)$$

$$-n(2\pi) - 60.4^\circ (\pi / 180^\circ) = -(2\pi / 14.38 \text{ m.}) L1$$

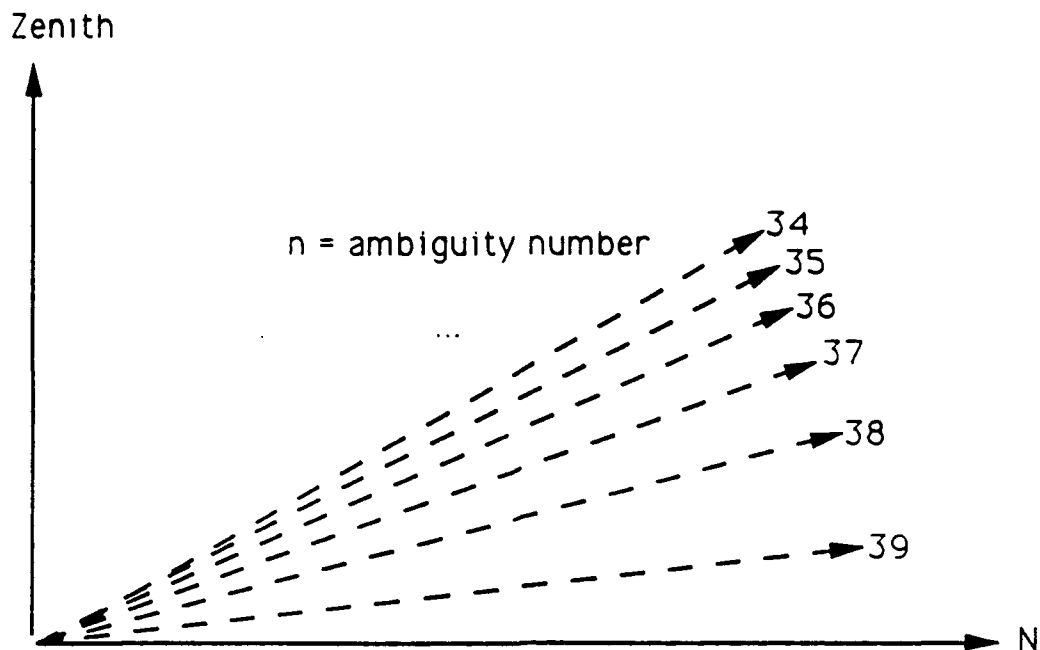
$$L2 = 15.1 \text{ m.} + n(14.38 \text{ m.})$$

$$EL = \arccos(L2 / 580 \text{ m.})$$

Note: Since $L2 \leq 580 \text{ m.}$, $n \leq 39$.

n	39	38	37	36	35	34
EL(°)	6.8	14.5	19.4	23.3	26.6	29.7

Figure 2.23a Angle-of-Arrival Elevation Calculations



n	39	38	37	36	35	34
EL(°)	6.8	14.5	19.4	23.3	26.6	29.7

Notes:

1. For the Ava,NY-FAU ground range of 2400km., Griffiths [33] indicates elevation angle-of-arrival ranges 6° – 12° for effective heights of 200–400km., respectively.
2. Since 720° phase shift corresponds to $19.4^{\circ} - 6.8^{\circ} = 12.6^{\circ}$ (or 219.9 mrad) of angle-of-arrival, 1° phase shift corresponds to .31 mrad.

Figure 2.23b Angle-of-Arrival Elevation Ambiguities

The technique was successfully tested in the laboratory prior to the field measurements.

In the field experiments, each receiver's 455 kHz IF was sampled at 52 KHz via a digital sampling oscilloscope. The memory depth of the oscilloscope was 16k per channel thus permitting two channels of ~24 sequential code sequences (~300 millisec) to be captured.

Two algorithms to convert to baseband were considered. One was developed by C. Rader [35] and the other by Waters and Jarrett at NRL [36-37]. In both laboratory tests and on field data, the NRL algorithm gave lower sidelobes and hence was used in the analysis.

Initial tests were made in March, 1991 at 24.5 MHz with a single, roof mounted antenna. In August, 1991 at 15.59 MHz, a field set of measurements was made using the E1-W1, 290m antenna separation shown in Figure 2.1. In these latter tests, for calibration purposes, a signal from one antenna was split equally into each receiver. The correlations peaks were approximately equal indicating equivalent processing gain in each channel.

In the March measurements, only a single correlation peak was observed whereas measurements made in August at 15.59 MHz yielded two or more peaks. These latter measurements were made in the daytime and the reports from Boulder indicated low solar activity and geomagnetic activity $A = 4$ and $K < 4$ indices. Examples of a single correlation peak for 1 code length and stacked 24 code lengths are shown in Figures 3.1 and 3.2. Note that over the 24 pulse sequence, no significant pulse amplitude variations are observed. The additional lower amplitude peak to the right is due to a processing sidelobe introduced by a small frequency mismatch between the receiver IF frequency and the sampling rate. This mismatch arose from the fact that at the receive site, separate and independent crystal controlled frequency synthesizers were used to generate the receiver 455 KHz and the oscilloscope 52 KHz sampling rates.

PN Data

24.5 Mhz, CW Mode, Roof Antenna, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00

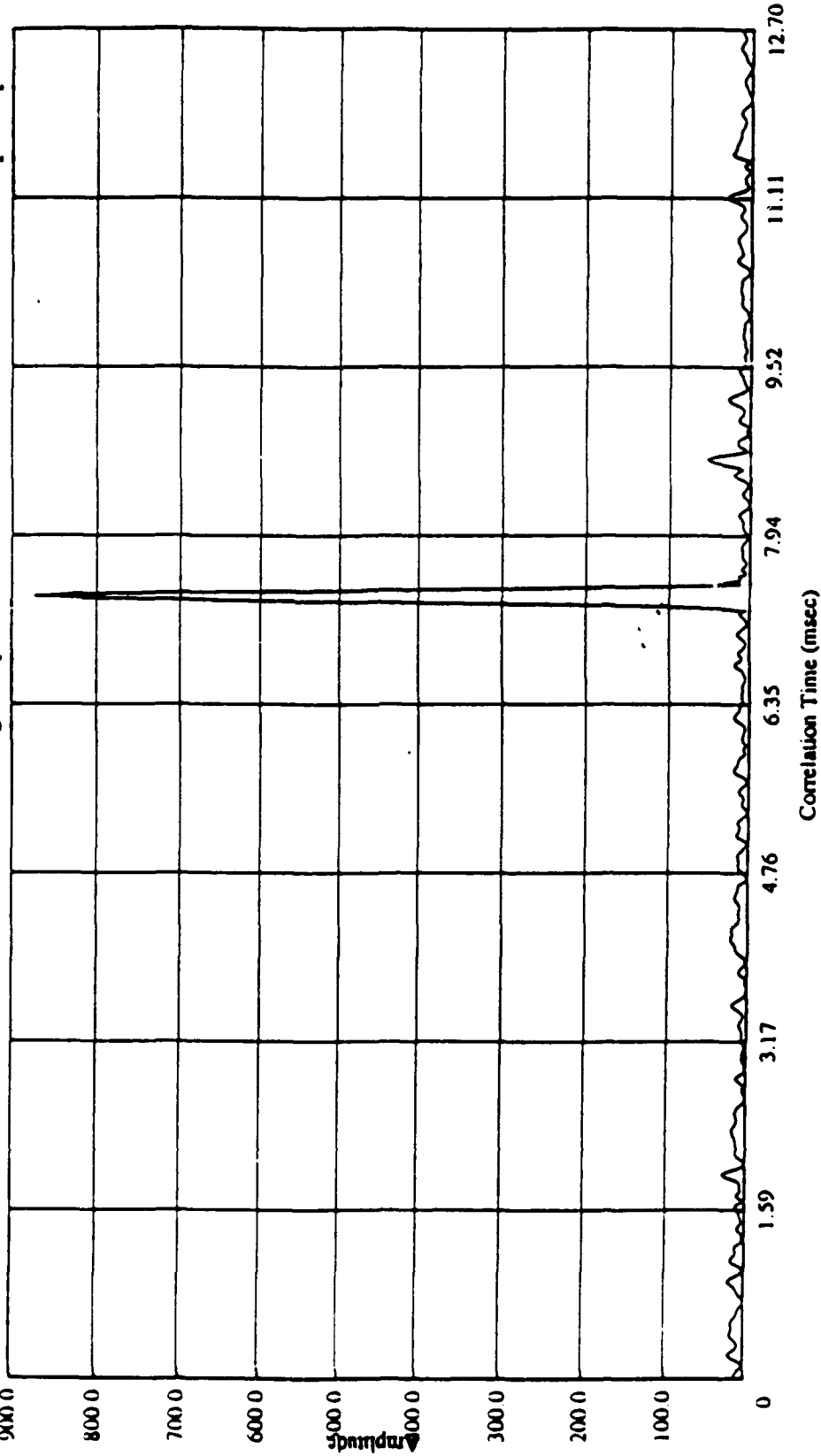
Sample Rate (Hz): 52000

Code Length: 127

Samples/Code: 660

Starting Sample: 0

Skip Sample: 0



Data File: mar01.05m

Fri Aug 2 16:07:38 1991

Figure 3.1

PN Data

24.5 Mhz, CW Mode, Roof Antenna, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00

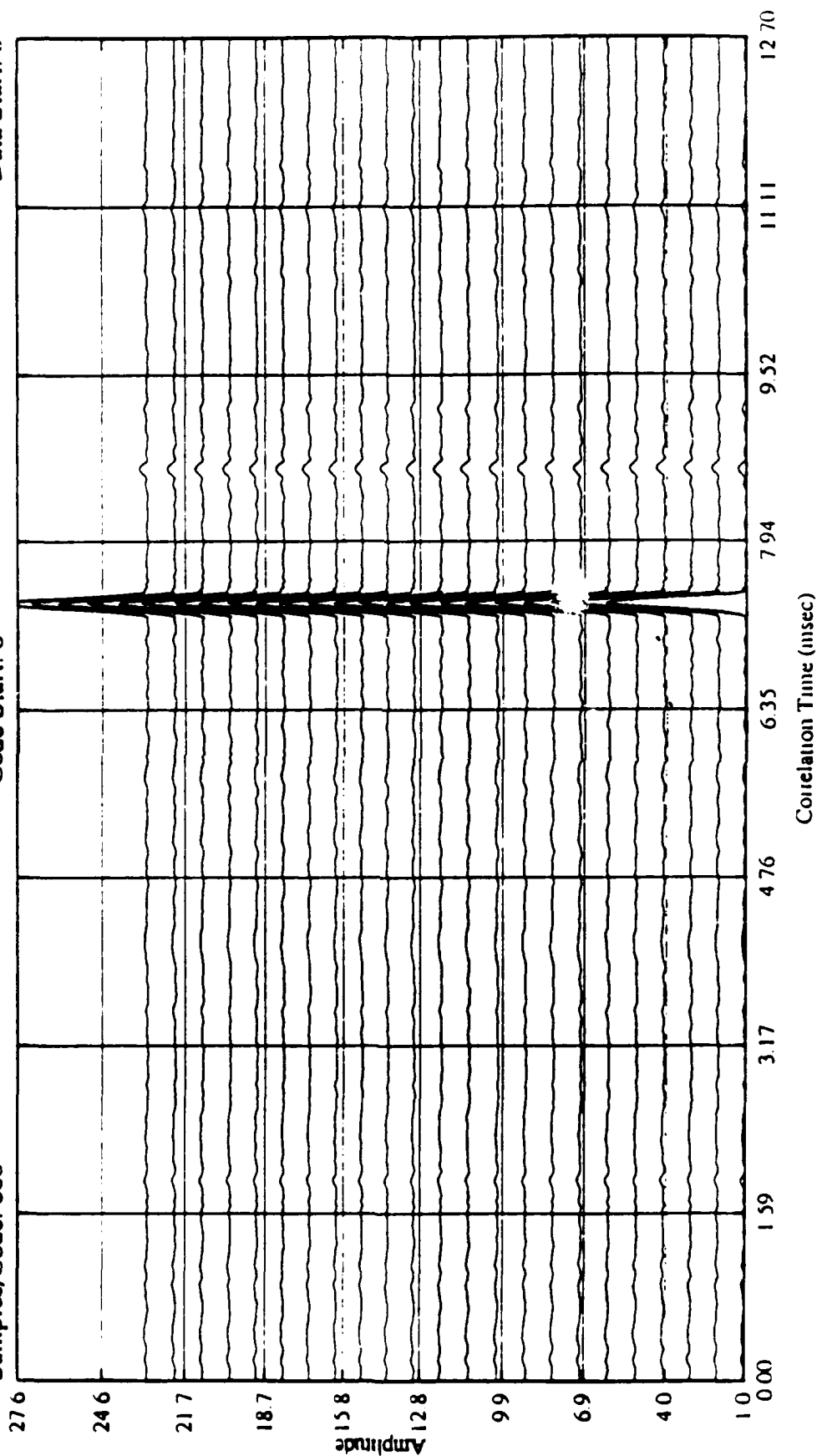
Sample Rate (Hz): 52000

Code Length: 127

Samples/Code: 660

Code Start: 0

Data Start: 0



Data File: c:\17.52

133

Data File: mar01.05m

Figure 3.2

Figures 3.3 and 3.4 illustrate the twin correlation peaks observed in the 15.59 MHz data. Again, small sidelobes appear at the right of the main peaks due to frequency synthesizer mismatches. Note that at one antenna, the first main peak is larger than the second whereas at the other antenna, the second is larger than the first. Between the major peaks, there is a clean separation of approximately 225 usec. This separation can be explained as multipath between the E-layer and F-layers. For an E-layer height of 100 km and F-layer height of 250 km., simple geometry yields a path difference of 235 usec.

Correlation snapshots at both antennas taken 3 minutes later are shown in Figures 3.5 and 3.6. Now at both antennas, the second correlation peak is higher than the first. During this short time interval the relative amplitude of each pulse has also changed significantly. If the phase of the signal at the first antenna would have been measured with the receiver set for CW, there would have been a significant phase jump as the dominant component changed from one peak to the other.

Figures 3.7 and 3.8 illustrate the signals captured simultaneously at the two receiving sites the previous day. In this case they are very dissimilar, each site having a different number of correlation peaks thus indicating the complex nature of the multipath over a short baseline distance of only 290 m.

These results should be taken to represent a proof of concept that direct IF sampling of PN codes can be used to resolve multipath. The oscilloscope has four channels and hence could be used to obtain simultaneous data from four antennas. However, two addition received would have to be acquired to perform these measurements.

From the sequential 24 pulse results, over a 300 millisec time interval, no significant amplitude changes were noticed. However, over a 3 minute interval, very significant amplitude changes were noticed occasionally. It would be of interest to obtain non-sequential data snapshots spanning the 3 minute period. This could be done with some additional electronic hardware to

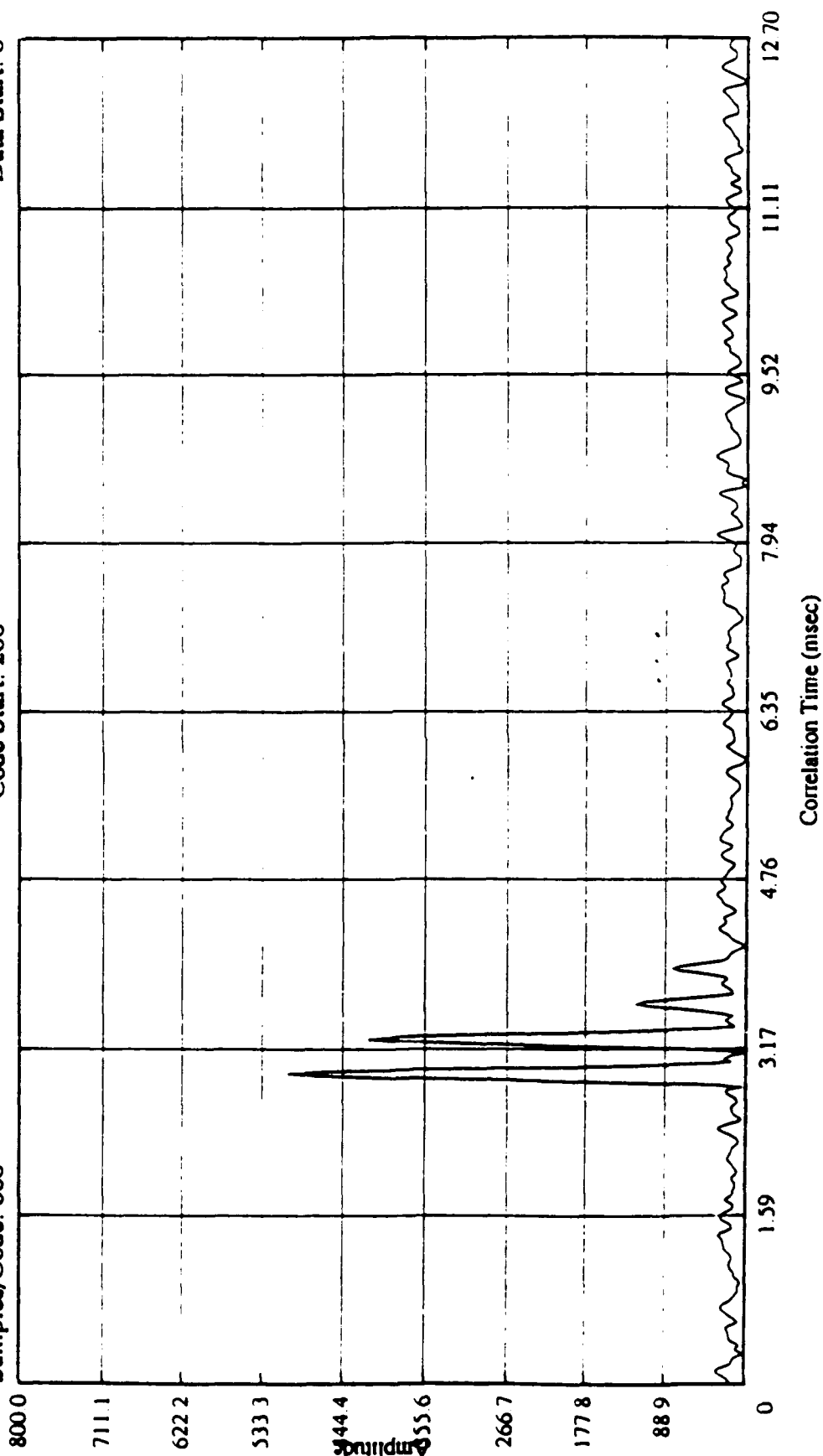
PN Data

15.59 Mhz, CW Mode, 0 dB Pad, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00
Samples/Code: 660

Sample Rate (Hz): 52000
Code Start: 200

Code Length: 127
Data Start: 0



Data File: c117.52

Time: 1991:8:15:10:46:56 EDT

Data File: aug15.03a

Figure 3.3

PN Data

15.59 Mhz, CW Mode, 0 dB Pad, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00

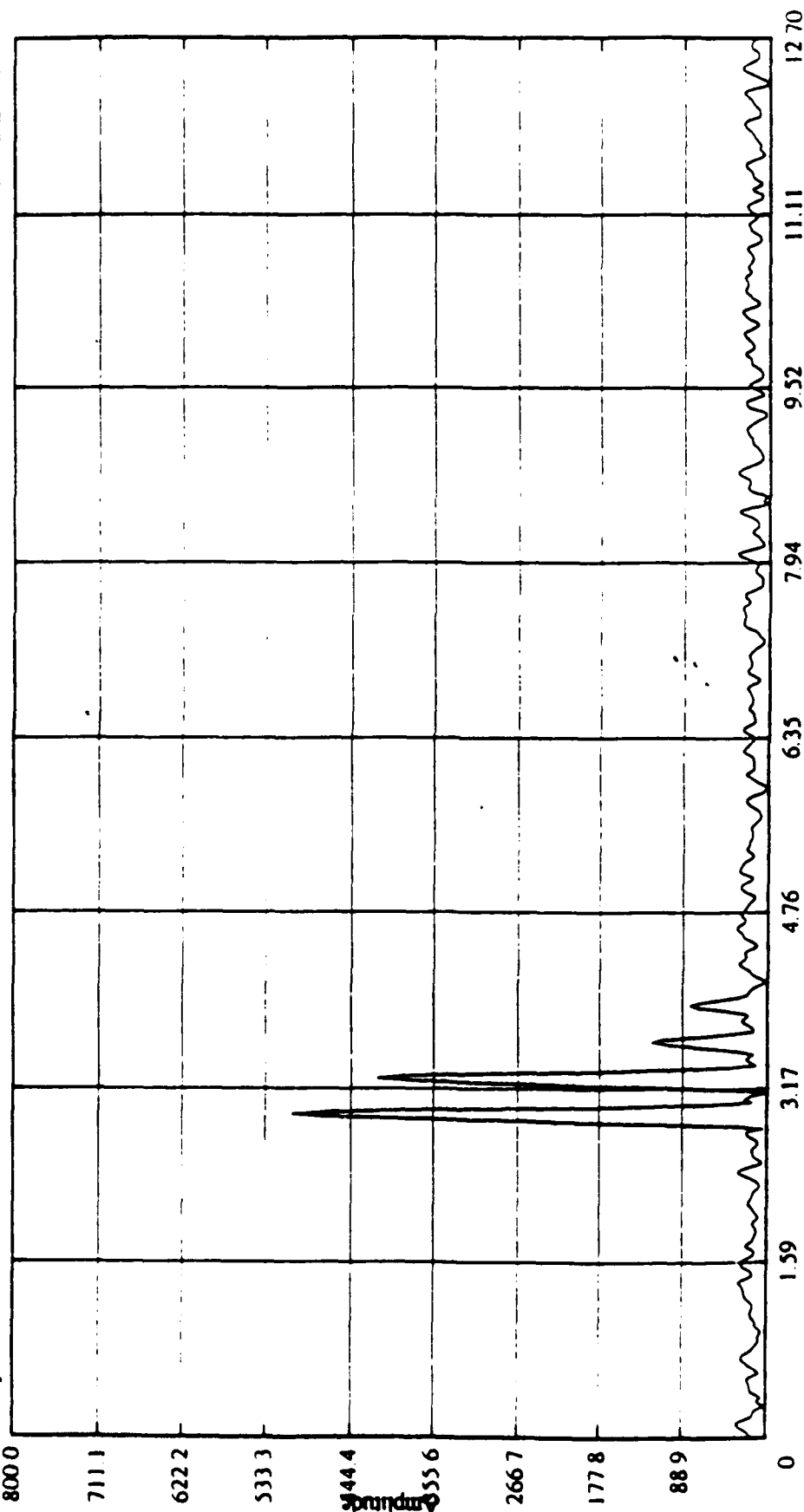
Sample Rate (Hz): 52000

Code Length: 127

Samples/Code: 660

Code Start: 200

Data Start: 0



Correlation Time (nsec)

Data File: c117.52

Time: 1991:8:15:10:46:56 EDT

Data File: aug15.03a

Figure 3.3

PN Data

15.59 Mhz, CW Mode, 0 dB Pad, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00

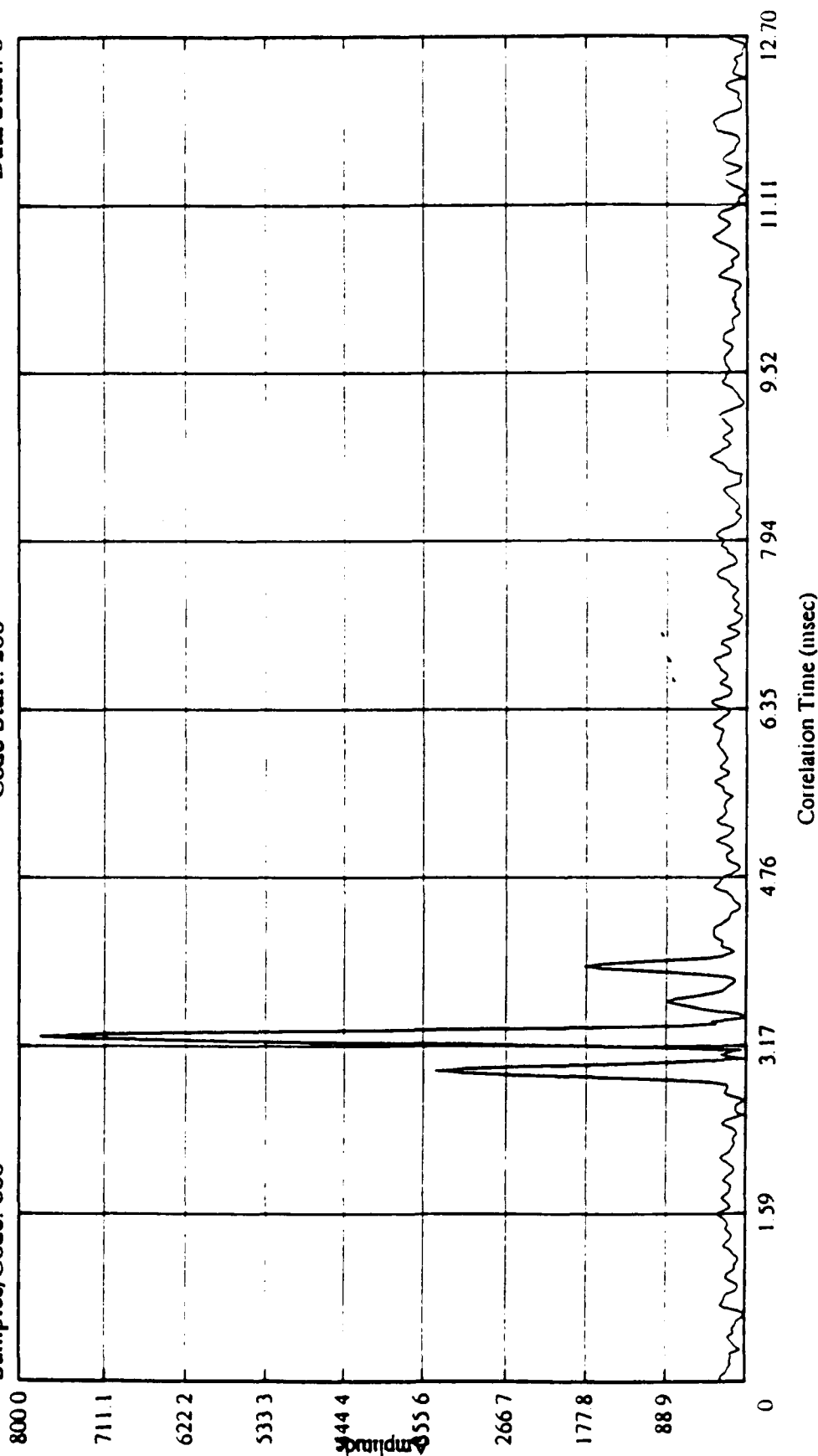
Sample Rate (Hz): 52000

Code Length: 127

Samples/Code: 660

Code Start: 200

Data Start: 0



Data File: c117.52

Time: 1991:8:15:10:46:56 EDT

Data File: aug15.03b

Figure 3.4

PN Data

15.59 Mhz, CW Mode, 0 dB Pad, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00

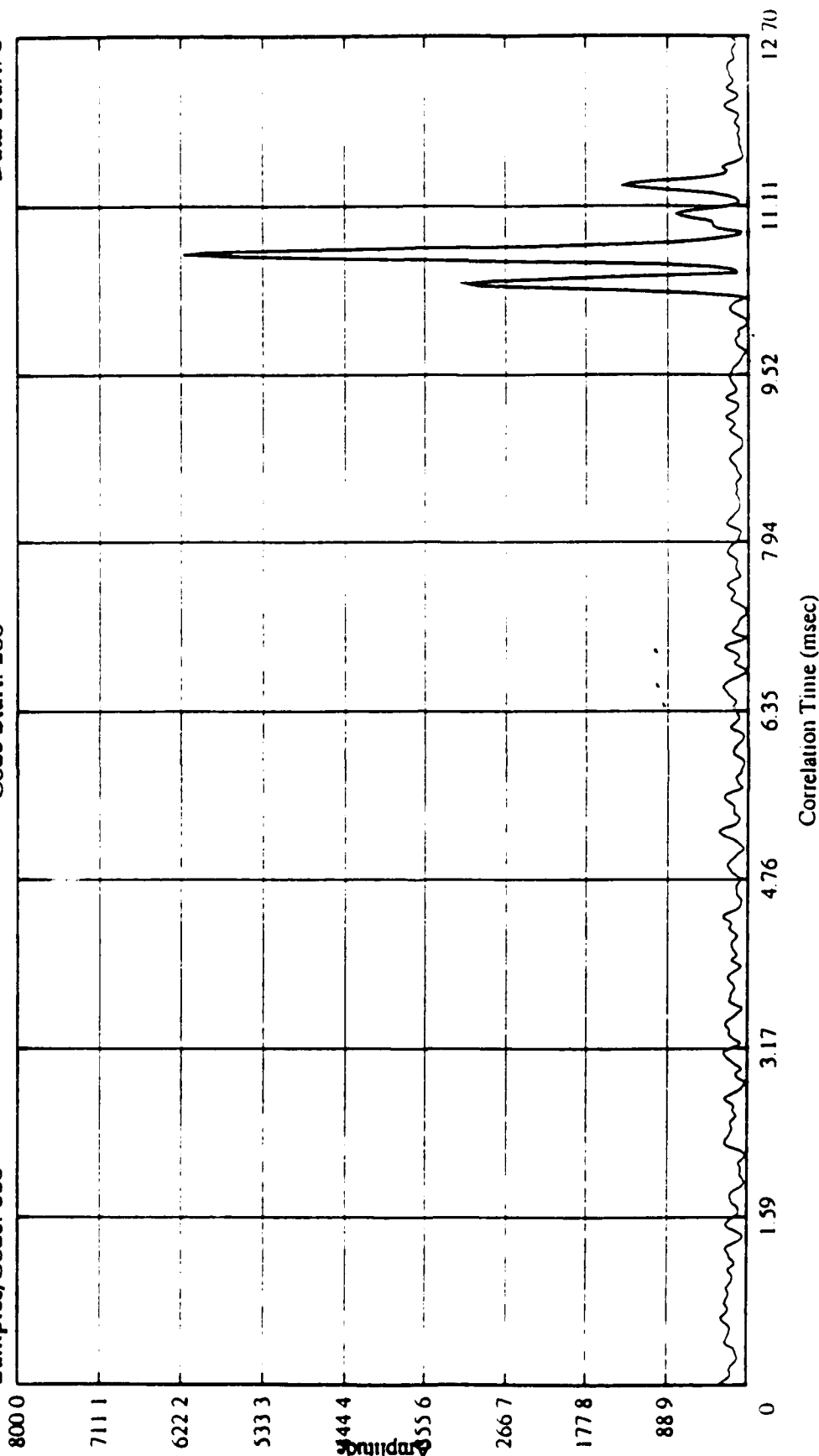
Samples/Code: 660

Sample Rate (Hz): 52000

Code Start: 200

Code Length: 127

Data Start: 0



Data File: c117.52

Time: 1991-8-15:10:49:44 EDT

Data File: aug15.04a

Figure 3.5

PN Data

15.59 Mhz, CW Mode, 0 dB Pad, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00

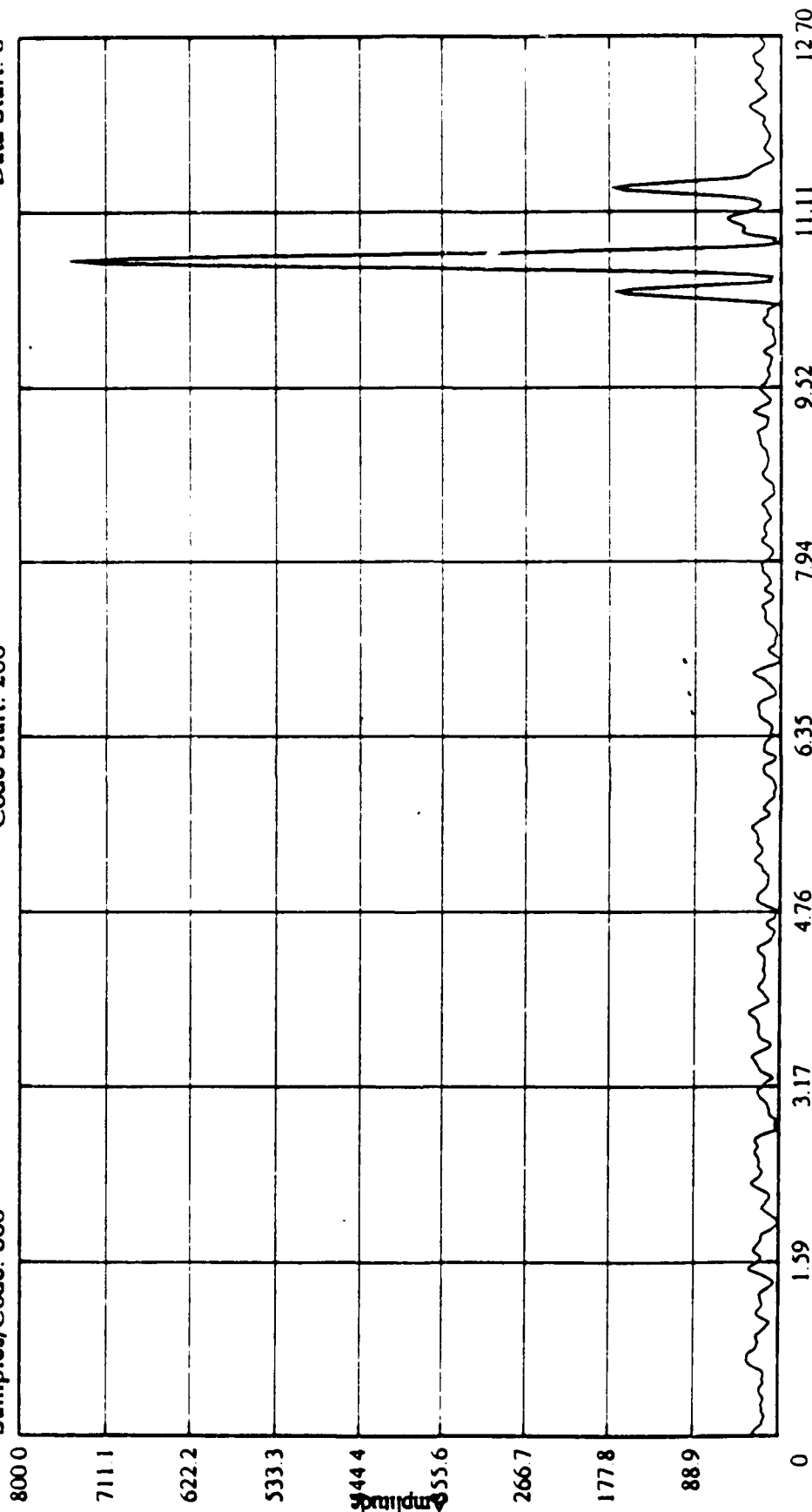
Sample Rate (Hz): 52000

Code Length: 127

Samples/Code: 660

Code Start: 200

Data Start: 0



Data File: c117.52

Time: 1991:8:15:10:49:44 EDT

Data File: aug15.04b

Figure 3.6

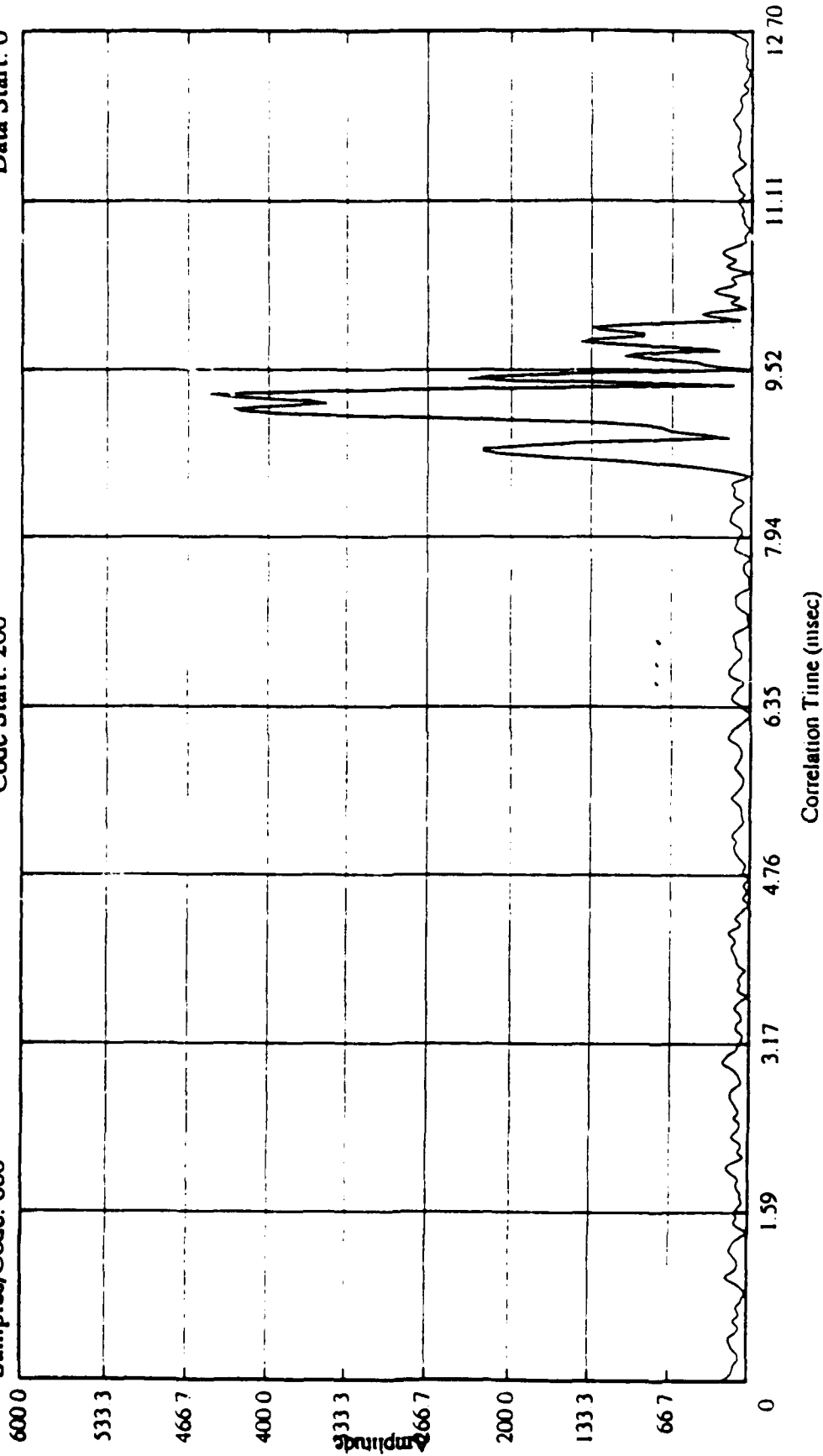
PN Data

15.59 Mhz, CW Mode, 0 dB Pad, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00
Samples/Code: 660

Sample Rate (Hz): 52000
Code Start: 200

Code Length: 127
Data Start: 0



Data File: c117.52

Time: 1991:8:14:12:46:10 EDT

Data File: aug14.12a

Figure 3.7

PN Data

15.59 Mhz, CW Mode, 0 dB Pad, 455 kHz IF into TEK 2214

Chip Length (usec): 100.00

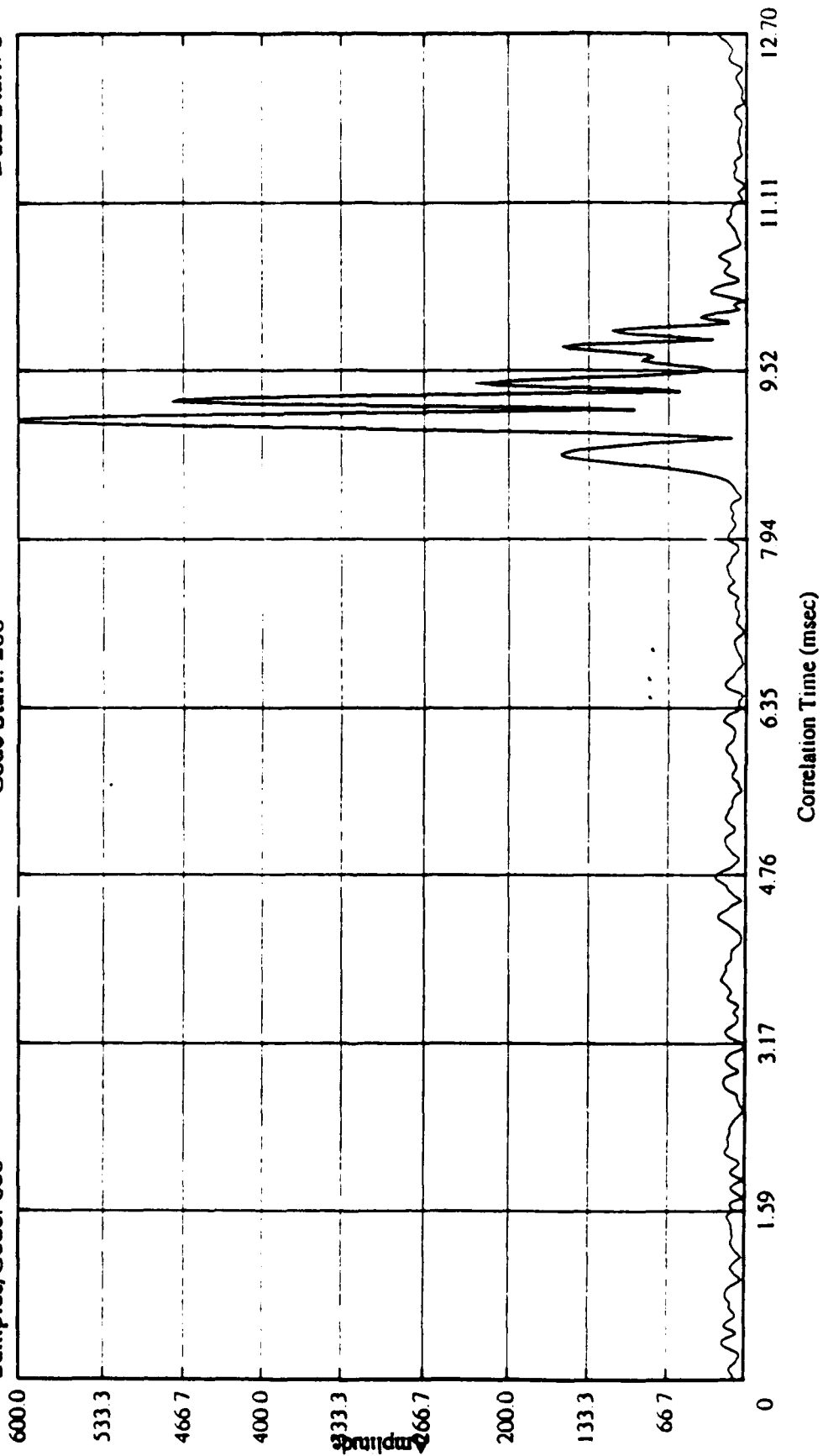
Sample Rate (Hz): 52000

Code Length: 127

Samples/Code: 660

Code Start: 200

Data Start: 0



Data File: c117.52

Time: 1991:8:14:12:46:10 EDT

Data File: aug14.12b

Figure 3.8

incorporate a burst control into the data sampling frequency synthesizer.

4.0 Conclusions

The means of azimuth and elevation angle-of-arrival were measured and found to be quite close to expected results from map information and published models. In addition, standard deviation of elevation was measured and yielded qualitatively expected results. However, the standard deviation of azimuth measurements did not follow the expected trend with larger baseline separation. Relative phase measurements between two widely spaced antennas produced results with periods of relative phase stability and other periods of sudden, large phase jumps. To investigate this further, mode separation techniques were incorporated into the experimental program.

Mode separation was done via the transmission of a 127 length PN code sequence. A new technique employed in these measurements was direct IF sampling and the conversion to baseband performed via software. The 455 kHz receiver IF amplitude was digitally sampled at 52 KHz with 8 bit resolution. Subsequent correlation to the original PN code sequence separated the signal cleanly into two components at 15.59 MHz and a single component at 24.5 MHz. The time separation of the two components at 15.59 MHz could be matched to multipath from the E- and F-layers. At 15 MHz, the relative amplitude of the 2 components could change significantly over short (minute) periods. For a receiver operating in a CW mode, a monitoring of the IF phase (and hence the carrier phase) would experience a significant jump when the dominant amplitude shift from one component to the other component. This effect is probably responsible for the sudden large phase changes seen in the original CW data. These jumps were not directly observed during the phase modulated measurements since these latter were only 300 millisecond snapshots.

It would be very helpful in future measurements to generate an ionospheric profile via a sounder. Locating the altitude of the significant ionospheric reflection layers could corroborate the observed path delays. Instrumental improvements could include

incorporating an improved time standard which could be used to lock to the transmitting waveforms and generate coherent receiver IF and sampling frequencies. Higher chip rate code sequences could be used to investigate multipath phenomena within a single layer. Some wideband probing has been reported in the literature [38]. It would also be of interest to transmit other codes sequences such as complementary codes. These latter codes have the property of having zero sidelobes.

Acknowledgements: The encouragement and support of Bert Weijers and Ray Cormier at RADC East is gratefully acknowledged. Without their help and patience in obtaining two reliable receivers, these measurements would never have taken place. We also want to acknowledge the operational help extended to us by Mr. Larry Tamburino and his staff at Ava, N.Y.

5.0 References

- [1] M. Skolnik, Introduction to Radar Systems. New York: McGraw-Hill, 1962.
- [2] J. Headrick and M. Skolnik, "Over-the-Horizon Radar in the HF Band," Proceedings of the IEEE, Vol. 62, June 1974, pp. 664-673.
- [3] S. Taheri and B. Steinberg, "A Measurement System and Analysis Procedure for Determining the Spatial Phase Structure Function of Ionospherically Reflected Waves," IEEE Trans on Antennas and Propagation, Vol. AP-27, July 1979, pp. 500-507.
- [4] B. Briggs, "The Physical Significance of the Correlation Ellipse in Ionospheric Drift Analysis," Radio Science, Vol. 11, October 1976, pp. 817-819.
- [5] P. Baker, "Spatial Correlation Measurements on One-Hop HF Radio Waves," IEEE Trans. on Antennas and Propagation, November 1971, pp. 793-794.
- [6] C. Rush, "Ionospheric Radio Propagation Models and

Predictions-A Mini-Review," IEEE Trans. on Antennas and Propagation, Vol. AP-34, September 1986.

[7] E. Altshuler, J. Morris, and B. Weijers, "Limitations Imposed by the Ionosphere on Over-the-Horizon Radars," RADC-TR-89-241, October 1989.

[8] K. Yeh and C. Liu, "Radio Wave Scintillations in the Ionosphere," Proceedings of the IEEE, Vol. 70, April 1982, pp. 324-360.

[9] H. Lai and J. Dyson, "The Determination of the Direction of Arrival of an Interference Field," Radio Science, Vol. 16, May-June 1981, pp. 365-376.

[10] M. Epstein, "The Effects of Polarization Rotation and Phase Delay with Frequency on Ionospherically Propagated Signals," IEEE Trans. on Antennas and Propagation, Vol. AP-16, September 1968.

[11] J. Bennett, "Doppler Shift Formulas for Waves in the Ionosphere," Radio Science, Vol. 11, July 1976, pp. 621-627.

[12] R. Crane, "Ionospheric Scintillation," Proceedings of the IEEE, Vol. 65, February 1977.

[13] D. Barrick, "FM/CW Radar Signals and Digital Processing," NOAA Technical Report ERL 283-WPL 26, July 1973.

[14] K. Davies, Ionospheric Radio Waves. Waltham, MA: Ginn, 1969.

[15] K. Davies, "Review of Recent Progress in Ionospheric Predictions," Radio Science, Vol. 16, November-December 1981, pp. 1407-1430.

[16] H. Rishbeth, "A Review of Ionospheric F Region Theory," Proc. of the IEEE, Vol. 55, January 1967, pp. 17-35.

[17] O. Villard, Jr., "The Ionospheric Sounder and its Place in the History of Radio Science," Radio Science, Vol. 11, November 1976, pp. 847-860.

- [18] M. Phillips, "Ground-Based Vertical-Incidence Ionograms," IEEE Trans. on Antennas and Propagation, Vol. AP-22m, November 1974, pp. 785-795.
- [19] "Special Issue on Topside Sounding and the Ionosphere," Proc. of the IEEE, Vol. 57, June 1969.
- [20] T. Damboldt, "Propagation Predictions for the HF Range by the Research Institute of the Deutsche Bundespost," Forschungsinstitut der Deutschen Bundespost Technical Report.
- [21] J. Aarons, "Global Morphology of Ionospheric Scintillations," Proc. of the IEEE, Vol. 70, April 1982, pp. 360-378.
- [22] G. Jacobs and T. Cohen, The Shortwave Propagation Handbook. Hicksville, NY: Cowan, 1979.
- [23] E.C. Jordan, Electromagnetic Waves and Radiating Systems, Englewood Cliffs, New Jersey; Prentice-Hall, 1964
- [24] IEEE Standard Definitions of Terms for Radio-Wave Propagation, ANSI/IEEE Standard 211-1977, 1977.
- [25] H. Booker, "The Use of Radio Stars to Study Irregular Refraction of Radio Waves in the Ionosphere," Proc. of the IRE, January 1958, pp. 298-314.
- [26] B. Steinberg, "A Proposed Approach for Increasing the Azimuthal Resolution of HF Radar," IEEE Trans. on Antennas and Propagation, Vol. AP-20, September 1972, pp. 613-618.
- [27] R. Crane, "Variance and Spectra of Angle-of-Arrival and Doppler Fluctuations Caused by Ionospheric Scintillations," Jour. of Geo. Research, Vol. 83, May 1978, pp. 2091-2102.
- [28] L. Humphrey, "Characteristic Ionospheric Multipath Phase Fluctuations," IEEE Trans. on Antennas and Propagation, March

1971, pp. 299-300.

[29] R. Kieburtz, "A Critique of Angle-of-Arrival Measurements by the Phase Difference Method," IEEE Trans. on Antennas and Propagation, July 1967, pp. 584-585.

[30] J. Barnum et al, "Measurement of HG Skywave Spatial and Temporal Coherence across a 10-km Receiving Antenna Aperture: System Development and Demonstration", Interim Technical Report 2178, SRI International, June 1989, p. 174.

[31] Coll,et.al., "Digital Detection of Coded-Pulse Ionosonde Signals," Proc. of the IEEE, February 1965, pp. 188-189.

[32] K. Dixon, Spread Spectrum Systems. New York: Wiley, 1976.

[33] D.M. Haines and B. Weijers. "Embedded HF Channel Probes/Sounders" 1985 IEEE Military Communications Conference (Milcom '85), October 1985, pp 253-259.

[34] J. Griffiths, Radio Wave Propagation and Antennas. Englewood Cliffs, New Jersey: Prentice-Hall, 1987.

[35] C.M. Rader. "A Simple Method for Sampling In-phase and Quadrature Components" IEEE Trans. Aerospace Electron. Syst. AES-20(6), 1984

[36] W.M. Waters, and B.R. Jarrett "Bandpass Signal Sampling and Coherent Detection" IEEE Trans. Aerospace Electron. Syst. AES-18(4), 1982, pp. 731-736

[37] W.M. Waters, G.J. Linde and B.R. Jarrett "Phase Accuracy Experiments with a Direct Sampling Coherent Detector" NRL Report 9182, March 1989, p. 10.

[38] L.S. Wagner et al, "Wideband probing of the Transauroral HF Channel: Solar Minimum", Radio Science, Vol 23, Number 4, July-August 1988, pp 555-568.

FINAL REPORT
1989 RESEARCH INITIATION GRANT
Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Conducted by the
Universal Energy Systems, Inc.

TITLE
A STUDY OF INTERACTING TUNNELING UNITS
WITH POSSIBLE APPLICATION TO HIGH TEMPERATURE SUPERCONDUCTORS

Principal Investigator: Michael W. Klein
Academic Rank: Professor
University:..... Worcester Polytechnic Institute
Department:..... Physics Department
Date:..... December 10, 1990.
Contract No:..... F49620-88-C-0053/SB5881-0378

This research was started during my 1989 Summer Faculty Research Fellowship at Rome Air Development Center, Hanscom Air Base, Boston, MA 01731. My local contact scientist at the Rome Air Development Center was ALFRED KAHAN.

1. ACKNOWLEDGEMENT

I thank the Air Force Office of Scientific Research for awarding me this Research Initiation Grant and for sponsoring this research. I also acknowledge Universal Energy Systems Inc. for their help in expediting the administrative aspects of this program; for their "instant information" supplied when needed in connection with this program.

My special thanks to Alfred Kahan of the Rome Air Development Center for having introduced me to a wealth of literature on high temperature superconductivity; for having spent a good deal of time to keep me informed of the newest literature in the field and for patiently explaining the structure of some of these high-temperature superconducting materials.

2. ABSTRACT

Most amorphous and glassy materials exhibit anomalies in their low temperature thermal properties. Similar anomalies were observed in a number of high temperature superconductors. It is believed that the anomalies in glasses arise from tunneling states in the solid, however, there is as yet no microscopic model for these anomalies. In this work we start with microscopic Hamiltonian for a set of multi-orientational tunneling (elastic) quadrupoles which interact with a quadrupolar interaction. From this Hamiltonian we obtain the free energy for low concentrations of tunneling units and derive the first completely microscopic theory for the low temperature properties of tunneling quadrupoles which are very similar to those observed in glasses.

We show that, contrary to the currently accepted theory on glasses, interactions play a very important role in the low energy excitations of tunneling quadrupoles. Our work suggests that current approaches to the theory of amorphous materials and glasses may have to be modified in order to make further progress in understanding them.

3. INTRODUCTION

Amorphous materials have a good deal of practical importance in technology and their understanding presents a serious challenge to the experimental as well as theoretical physicist. There has been a good deal of recent interest in the study of the properties of amorphous materials.[1,2] This interest arises because experiments on glasses and amorphous materials exhibit low temperature anomalies in their specific heat,[3-4] thermal conductivity, thermal expansion, dielectric relaxation and sound propagation and absorption. The early experiments concentrated mostly on amorphous insulators. More recent[5] ones found that the glassy anomalies are also present in amorphous metals.

It is quite remarkable that almost all amorphous materials show very similar low temperature anomalies. To explain these anomalies it was proposed that the "glassy"[6] system is made up of two-level tunneling states (TLS) [3,4] with a random distribution of barriers between the two wells. An ad-hoc assumed phenomenological constant density of states gives many of the experimentally observed properties of glasses, and thus presents a good deal of experimental support for the existence of tunneling states in glasses.[1,2] However, all attempts to obtain a microscopic description for the tunneling states or for the constant density of states have had very limited success.

In addition to most amorphous materials, there are a number of other systems that exhibit "glasslike" properties at low temperatures. Dilute concentrations of tunneling electric dipoles[7-8], (for example Li^+ in KCl) or (elastic) strain dipoles[9] (for example CN^- dissolved in KCl) also show low temperature glasslike anomalies. In fact there are whole classes of impurities which when dissolved in alkali halides show glass-

like properties of one form or another.[9-14]

More recently there has been a good deal of evidence in the literature for glassy properties in a number of high temperature superconductors. Evidence for possible glassy properties have been seen in the sound velocity[15-19], in the thermal conductivity[20-21] and in the specific heat.[21] It is the observation of the glassy properties of some of the high temperature superconductors that has motivated this research.

In an attempt to identify the origin of the low-energy excitations in tunneling units, I studied[22] the microscopic properties of very dilute two-orientational Ising-model tunneling-dipoles which interact via a $1/r^3$ interaction. I derived a set of microscopic relations for the thermal properties of the tunneling dipoles. From the microscopic derivations I obtained a constant density of excitation energies E for low E in agreement with experimental results observed in glasses. Later on[23,24] I treated multiorientational tunneling dipoles (for example Li^+ in KCl) and obtained the specific heat, the thermal expansion and the complex dielectric susceptibility[24] for the tunneling dipoles. Again these results are consistent with those observed in glasses. Thus the tunneling dipole problem[24] serves as a first microscopic model for one particular glass system.

In spite of the successful derivation of glasslike behavior[24] for the tunneling dipoles, the derivations are not expected to shed much understanding on real glasses which do not in general exhibit large local dipole moments. They do, however, show low energy excitations which affect the elastic constants and the sound velocity. Such effects are believed to be primarily strain dominated and were observed in high T_c superconductors.[15-19]

4. OBJECTIVES OF THE RESEARCH

The objective of the research was to formulate a microscopic approach to derive the thermal properties of dilute interacting tunneling (elastic) strain quadrupoles and to use this formulation to obtain expressions for: (a) the density of states (excitation energies); (b) the specific heat; (c) the thermal conductivity; (d) the sound absorption and propagation and (e) the dielectric properties. A further objective was to study the glasslike properties of high T_c superconductors. It was hoped that the understanding of the glasslike properties of tunneling quadrupoles will also help to understand some of the glasslike properties of high temperature superconductors.

5. RESEARCH ACCOMPLISHED

In this section we briefly describe the research accomplished through the support of this grant.

As outlined in the introduction, there is currently no microscopic theory which explains the universal properties of amorphous materials or the glasslike properties of high temperature superconductors. With the help of work which I started during my Summer Faculty Research Fellowship at Rome Air Development Center (RADC) I have been able to develop a microscopic theory for dilute tunneling quadrupoles randomly distributed in crystalline hosts. This microscopic theory gives glasslike properties for (elastic) tunneling strain units in a solid from fundamental considerations.

First we formulated the problem of interacting (elastic) tunneling quadrupoles randomly distributed in a nonpolar host. We allowed the tunneling quadrupoles to interact via a strain dominated interaction. We obtained low energy excitations similar to those observed in glasses. A preliminary report on this work was presented at the American Physical Society Meeting, March 1990, Los Angeles, CA.[25] A copy of this publication is included as Appendix A to this report.

Next we examined the low temperature thermal conductivity and found this also to be very similar to that found in glasses. A preliminary report on this work was presented at the American Physical Society Meeting, March 1990, Los Angeles, CA.[26] A copy of this publication is included in Appendix B to this report.

Next we derived the low temperature specific heat for the tunneling quadrupoles. We showed that our temperature dependent specific heat predictions are in very good agreement with low T experiments[27] on 320

ppm CN^- in KBr and give a near-neighbor CN^- - CN^- interaction $K_0 \leq 100$ K. Previous calculations, using the model developed for dipoles[22] gave a value of K_0 much larger than both, 100 K and the estimated correct value.[27] This theory also explains the low energy excitations observed in rotary-echo experiments[28] from CN^- pairs in KBr. Thus the derivations for the quadrupoles explain the two experiment so far available. However the theory predicts all the thermal properties of the quadrupoles, many of which have not yet been measured. It is hoped that this theory will stimulate a series of new experiments to compare with the predictions of the theory. Some suggested measurements are: (a) the concentration dependence of the specific heat for low concentrations and low temperatures; (b) the thermal conductivity; (c) the temperature and frequency dependent dielectric constant.

Upon deriving our temperature-dependent specific heat results they were presented at the International Conference on Relaxation in Solids at Crete, Greece, June 1/-29, 1990. These results were written up and were accepted for publication by the Journal for Noncrystalline Solids and will appear in April 1991.[29] A copy of this paper is enclosed as Appendix C of this report.

In further derivations we showed that there are important differences in our approach for the tunneling quadrupoles problem and that of the accepted TLS model. These differences are: in the TLS^{3,4} the low energy excitations arise from isolated TLS with an assumed random distribution of potential barriers, in our derivations they arise from the strongly interacting tunneling quadrupoles; in the TLS model the tunneling matrix element Δ must be very small to contribute at low T, in our derivation Δ may be large for sufficiently strong interactions; in the TLS model the

constant density of states is assumed without justification, in our work the constant density and other glasslike⁵ properties are derived from fundamental principles; the TLS model can not derive the microscopic dielectric constant or determine whether the phonons couple to the TLS via the tunneling matrix element Δ or the anisotropy energy, while in our work these quantities are directly obtained from a microscopic derivation. Our work thus gives the first completely microscopic model for the glassy state of tunneling quadrupoles without using the mean field approximation. Mean field approximation gave microscopic models for higher concentrations than the ones treated here.[30,31]

We also showed that there are significant differences between the results obtained for the tunneling dipoles[24] and that for the tunneling quadrupoles derived here. For the 2^n -orientational tunneling dipoles the partition function factors into products of Ising partition functions, this is not so for the tunneling quadrupoles. The microscopic expressions for the dielectric constant, specific heat and thermal conductivity are different for the two cases. An example of this difference is exhibited by the calculation of the CN^- near-neighbor interaction K_0 further on in this letter. Using the theory for the tunneling quadrupoles gives $K_0 \approx 100$ K, whereas using the dipole calculations[22] with a factorization of the partition function gives [27] $K_0 \approx 450$ K. The latter is much greater than what is believed to be the experimental value.[27]

The results of these derivations were submitted and accepted for publication in the Physical Review Letters, published on December 10, 1990.[32] A copy of the manuscript of this publication is also enclosed as Appendix D to this report.

Two other articles on the detailed derivation of these results have been submitted to the Physical Review under the titles: "Glasslike excitations from dilute interacting tunneling units I: Four orientational tunneling quadrupoles.[33] Finally and "thermal conductivity of dilute tunneling quadrupole: Are the two types of tunneling unit.[34] Copies of both articles is enclosed in the Appendix E and Appendix F to this report.

The glassy state at higher concentrations. The work outlined above obtains a solution of the dilute quadrupole problem for very low concentrations of tunneling quadrupoles distribute in a nonpolar medium. The derivations give low energy excitations with a constant density of states. The derivations arise from a virial expansion of the free energy in the impurity concentration c and therefore is valid only for very low c . The low energy excitations arise from the strongly interacting tunneling quadrupoles.

What happens at higher concentrations when the virial expansion is no longer valid? In such a case one has to resort to mean field or other approximations to examine the glassy state. In order to do so we studied the low energy excitation spectrum of dilute concentrations of interacting tunneling quadrupoles randomly distributed in a non-polar medium the mean-field approximation. In particular we consider the case of six orientational tunneling quadrupoles with a r^{-3} (elastic) interaction. To examine the nature of the low energy excitations for higher concentrations we derive the free energy and the specific heat for six orientational tunneling quadrupoles in the mean field approximation. The internal field is a random variable and for relatively low concentrations has a Lorentzian distribution. We find that the low energy density of states is a constant and that the low energy excitations arise from the large internal fields,

i. e. strongly interacting tunneling quadrupoles just like for the case of the very low concentration limit discussed above. This work, entitled "Low energy excitations from interacting tunneling units in the mean-field approximation" was presented at the International Conference of Relaxation of Complex Systems, Crete, Greece, June 17-29 1990; it was also submitted to the Journal of Noncrystalline Solids and will be published in April 1991.[35] A copy of this manuscript is included as Appendix G to this report.

We summarize the work accomplished under this contract. A good deal of progress was made in the understanding of the glasslike properties of the tunneling units at low temperatures and of the nature of the glassy state. Similar progress was made in the understanding of the specific heat, thermal expansion, sound absorption and propagation and dielectric properties of the tunneling quadrupoles. We have derived the first microscopic theory for the glasslike state without using the mean field approximation. Our approach is completely different from the currently accepted theories on glasses and is quite promising that it will lead to a better understanding of the microscopics of amorphous systems; this has eluded physicists for the past eighteen years in spite of considerable effort theoretical effort.

The approaches mentioned above for the tunneling dipole and tunneling quadrupole problem can be quite likely also applied to the glasslike properties of high T_c superconductors. This study will be undertaken in the near future.

6. RECOMMENDED FURTHER WORK

The work accomplished in this proposal has been highly successful in furthering our understanding the current state of the low energy excitations in randomly distributed tunneling units. The work accomplished so far seems to point the way to further research in this area. It has already helped clear up some of the reasons for the low temperature anomalies in amorphous materials and glasses. A good approach to attack the general problem seems to be a combination of low concentration solutions with mean-field approaches at higher concentrations. The understanding of this problem could also help improve the possible prediction of new amorphous materials with various desirable properties.

Next we would like to address ourselves to the glassy properties of high T_c superconductors. These glassy properties have an important effect on the pinning of the flux lines, on the critical current densities and on other physical properties in high T_c superconductors. It is therefore very important to understand their glassy nature. It is suggested that further work should be done using some of the methods derived in the enclosed papers as a guide to understand the glassy properties of high T_c superconductors.

7. REFERENCES

- [1]. Amorphous Solids, edited by W. A. Phillips, (Springer Verlag, Berlin, 1981).
- [2]. S. Hunklinger and W. Arnold, in Physical Acoustics Vol. 12, edited by W. P. Mason and R. N. Thurston, (Academic Press, New York, 1976), p.155.
- [3]. P. W. Anderson, B. I. Halperin and C. Varma, Phil. Mag. 25, 1(1972).
- [4]. W. A. Phillips, J. Low Temp. Phys. 7, 351(1972).
- [5]. S. Hunklinger and J. Raychaudhuri, in Progress in Low Temperature Physics Vol 9, edited by D. F. Brewer (North Holland , Amsterdam-New-York, 1986) and references therein.
- [6]. Materials which have properties similar to amorphous solids and glasses, but are not usually considered to be glasses will be denoted as "glasslike".
- [7]. R. C. Potter and A. C. Anderson, Phys. Rev. B24, 4826(1981).
- [8]. R. C. Potter and A. C. Anderson, Phys. Rev. B24, 677(1981).
- [9]. D. Moy, R. C. Potter and A. C. Anderson, J. Low Temp. Phys. 52, 115(1983).
- [10]. W. Kanzig, H. R. Hart and S. Roberts, Phys. Rev. Lett. 13, 543(1964).
- [11]. A. T. Fiory, Phys. Rev. B4, 614(1971).
- [12]. M. W. Klein, C. Held, and E. Zuroff, Phys. Rev. B13, 3576(1976).
- [13]. M. W. Klein, B. Fischer, A. C. Anderson and P. J. Anthony, Phys. Rev. B17 4997(1978).
- [14]. U. T. Hoechli, Phys. Rev. Lett. 48, 1494(1982).
- [15]. M. J. McKenna, A. Hikata, J. Takeuchi, C. Elbaum, R. Kershaw and A. Wold, Phys Rev. Lett. 62 1556(1989)
- [16]. B. Golding, N. O. Birge, W. H. Haemmerle, R. J. Cava and E. Rietman Phys. Rev. B36, 5606(1987).

- [17]. A. Jezowski, J. Klamut, R. Horyn and K. Rogacki, Superc. Sci. Tech. 2, 96(1988)
- [18]. S. D. Peacor and C. Uher, Phys. Rev. B39, 11559(1989).
- [19]. S. J. Burns, A. Goyal, and P. D. Funkenbusch, Phys. Rev. B39, 11457, 1989.
- [20]. P. Esquinazi, C. Duran, C. Fainstein and M. Nunez Regueiro, Phys. Rev B37, 545(1988).
- [21]. For a recent review article on the specific heat see S. E. Stupp and D. M. Gindsberg, Physica C158, 299(1989).
- [22]. M. W. Klein, Phys. Rev. B29, 5825(1984); B31, 2528(1985).
- [23]. M. W. Klein, Phys. Rev. B35, 1397(1987).
- [24]. M. W. Klein, Phys. Rev. B40, 1918(1989).
- [25]. Published in Bulletin of the American Physical Society, Volume 35, Page 544, 1990.
- [26] Published in Bulletin of the American Physical Society, Volume 35, Page 545, 1990.
- [27]. J. N. Dobbs, M. C. Foote and A. C. Anderson, Phys. Rev. B33, 4178(1986).
- [28]. G. Baier, M. v. Schickfus and C. Enns, Europhys. Lett. 8 487(1989).
- [29]. M. W. Klein, Journal of Noncrystalline Solids, to be published in April 1991.
- [30] Mean field calculations were used to treat highly concentrated mixed crystals of KCl-KCN⁻ for c much beyond the validity of our low concentration work. See J. P. Sethna and K. Chow, Phase Transition 5, 317(1985).
- [31] M. Meissner, W. Knaak, J. P. Sethna, K. Chow, J. J. DeYoreo and R. O. Pohl, Phys. Rev. B32, 6091(1985). However mean field results are

problematic for competing interactions as was pointed out by I. Kanter and H. Sompolinsky, Phys. Rev. B33, 2073(1986).

[32] M. W. Klein, Phys. Rev. Lett.65, 3017(1990).

[33] M. W. Klein, Submitted to the Physical Review

[34] A. Galasso and M. W. Klein, Submitted for publication.

[35] P. Nielaba and M. W. Klein, Journ. Noncrystalline Solids, April 1991.

PUBLICATIONS ARISING FROM THIS GRANT

1. Microscopic Theory of Dilute Quadrupole Glasses. Physical Review Letters, Volume 65, page 3017, December 1990.
2. A solvable Microscopic Model for very dilute Quadrupole Glasses, to be published in Journal of Noncrystalline Solids, April 1990.
3. Low Energy Excitations from Tunneling Units in the Mean Field Approximation. With Peter Nielaba. To be Published in Journal of Noncrystalline Solids, April 1990.
4. Thermal Conductivity of Tunneling Quadrupoles: Are there tow types of tunneling units? With A. Galasso. Submitted for publication.
5. Glasslike excitations from dilute interacting tunneling units I: Four orientational tunneling quadrupoles. Submitted for publication.

PRESENTATIONS AT SCIENTIFIC CONFERENCES

1. Microscopic theory of dilute quadrupole glasses, presented at the 1990 March APS Meeting, Los Angeles, California, Bull. Am. Phys. Soc. Vol 35, Page 544.
2. Low temperature thermal conductivity in dilute interacting tunneling quadrupoles, with A. Galasso, Presented at the 1990 March APS Meeting, Los Angeles, California, Bull. Am. Phys. Soc. Vol 35, Page 545.
3. A solvable microscopic model for very dilute quadrupole glasses. **Invited paper** at the International Conference on Relaxation Phenomena in Solids, Crete, Greece, June 17-29, 1990
4. Publication # 3 above was presented at the International Conference on Relaxation in Complex Systems, Crete, Greece, June 17-29 1990.

Appendices may be obtained
from the author
or from UES

FINAL REPORT

REDUCED BANDWIDTH BINARY PHASE-ONLY FILTERS

Submitted to:

**Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432**

Prepared by:

**William L. Kuriger
School of Electrical Engineering
and Computer Science
University of Oklahoma
Norman OK 73019**

28 June 1991

ABSTRACT

REDUCED BANDWIDTH BINARY PHASE-ONLY FILTERS

A binary phase-only optical correlator, using an inexpensive liquid-crystal television as the spatial light modulator for the filter function, was developed and tested. A number of simulations, and a small number of experimental runs, investigated the effects of reducing the bandwidth of the filter. The low-pass effect allows a trade of specificity for tolerance, and binary phase-only correlation intensity peaks can be made to look much those obtained for a matched filter. A high-pass prefiltering operation can increase specificity, at a cost of reduced tolerance to scale changes and rotations. The quality of the experimental correlator performance obtained was not good, but was adequate for demonstration purposes, and could be improved with further development.

REDUCED BANDWIDTH BINARY PHASE-ONLY FILTERS

INTRODUCTION

The purposes of this research program were

1. To set up an experimental correlator system to verify simulation results and provide guidance into effects that need to be included in simulations.
2. To investigate, both theoretically and experimentally, the advantages and disadvantages of operating the phase-only filter at reduced spatial bandwidths (low-pass, high-pass, and bandpass).

To accomplish these purposes, procedures followed included running a number of simulations on diverse patterns, designing a test bed correlator based on an inexpensive liquid crystal television as the spatial light modulator (the correlator filter), and testing the correlator system, with special attention given to the effect of reducing spatial bandwidth.

PRELIMINARY DISCUSSION: BASIC CONCEPTS

The use of binary phase-only filters in optical correlators was suggested and popularized by Horner and co-workers [1,2,3]. Work in image processing had previously shown that most of the information content of images resides in phase information, with very little attributable to amplitude information [4]. Details of the theory behind the use of binary phase-only filters in correlators were subsequently mapped out by many researchers.

A very simplified overview of the concepts included in this report can be obtained by considering a one-dimensional pulse, such as that shown in Figure 1. The pulse depicted is of unit height and is 9 pixels wide (in a field of 128 pixels). The Fourier transform of this pulse is the familiar $\sin(x)/x$ waveform shown in Figure 2. Note that much of the area under the curve is in the central lobe, the low-frequency Fourier components, with diminishing amplitudes at higher spatial frequencies. Fourier components are normally complex, but for this special case of a symmetric waveform centered about the coordinate origin, all amplitudes have phase angles of 0 or π radians. The application of a binary phase-only filter (BPOF) to this spectrum is precisely equivalent to simply taking the magnitude. The magnitude of the transform of the pulse waveform is shown in Figure 3. Retransforming the filtered spectrum, that is, taking the Fourier transform of the magnitude of the Fourier transform of the rectangular pulse, results in the correlation waveform shown in Figure 4. The correlation shows a narrow high peak at the location of the original pulse, and some much weaker sidelobes are in evidence. This is behavior that is typical of BPOF correlators.

The high peak results from the addition of in-phase components, so if the filter blocked further-out lobes from contributing to the transform summation, the observed peak would become lower and broader. On the other hand, it would then be less sensitive to any differences in scale or angle rotations. Reducing the bandwidth of a phase-only filter allows a correlator to be designed with any desired tradeoff between specificity and tolerance.

SIMULATIONS

The liquid crystal tv used as a spatial light modulator filter does not have sufficient resolution to use images of any complexity as test patterns. Consequently, all test patterns used were relatively simple geometric shapes, and a binary amplitude pattern was used (no gray scale). Three patterns were used in the investigations: A simple square (Figure 5), a stylized version of an airplane (Figure 6), and text spelling out "OKLAHOMA SOONERS" (Figure 7).

The Fourier transform magnitude of the 39x39 pixel square of Figure 5 is depicted in Figure 8. Correlations of this square with filtered versions of itself are given in Figures 9 through 14. All correlations are plots of intensity, not amplitude. A matched filter (MF) is used in the correlation shown in Figure 9, while a binary phase-only filter (BPOF) is used in the correlation depicted in Figure 10. Figures 11 and 12 show the effect of reducing the bandwidth of the BPOF. Figure 11's filter passed only the main lobe of the Fourier transform, while the filter used in Figure 12 passed the main lobe plus the next ones (zero order and first order terms). Note the similarity in appearance and peak value to the matched filter, suggesting that reduced-bandwidth BPOFs can be used as approximations to matched filters, at least for images that are not too complex. Figure 13 shows the correlation obtained for a high-pass filter. All but the central lobe is used for this filter. As expected, a correlation involving a high-pass filtering operation results in a very sharp peak, with the concomitant increase in sensitivity to rotations or scale changes. Figure 14 shows a bandpass filter correlation. Only the 2nd, 3rd, and 4th lobes are used in constructing this filter. Note that it has a much reduced correlation peak, suggesting that its use would be restricted to distinguishing nearly identical images.

The Fourier transform magnitude for the plane of Figure 6 is shown in Figure 15. Figure 16 shows correlations for the plane image using a BPOF, a CPOF (continuous phase-only filter), and an MF. The peak to sideband ratio (PSR) is the ratio of the correlation peak value to the sum of pixel values that are less than half of the peak value. Figure 17 is the same three correlations, but with added noise this time. The added noise was of sufficient amplitude to change, on the average, every 4th pixel value. A signal to noise ratio (SNR) was calculated as the ratio of the correlation peak intensity to the average noise intensity. As expected, the

matched filter performs best for this measure, since it is optimized to do just that. The BPOF performance is considerably degraded, but it does still clearly show a correlation peak. Figure 18 shows correlations, with and without added noise (of the same 0.25 watt per pixel level as before), for the plane image using a reduced bandwidth BPOF. Only the central 4 lobes of the Fourier transform are used for this simulation. A comparison of Figure 18(b) with Figure 17(a) shows that reducing the filter bandwidth improves the SNR by a factor of 2.15 in this case.

The last set of simulations was based on an image consisting of the letters OKLAHOMA SOONERS, previously shown in Figure 7. The interesting aspect of this particular simulation is that correlation is used to distinguish symbols whose properties are quite similar. The Fourier transform magnitude for OKLAHOMA SOONERS is shown in Figure 19. Figure 20 depicts Fourier transform magnitudes for the letters O and S by themselves, and Figures 21 and 22 show the central columns of the 3-dimensional graphs of Figure 20. The transforms of these letters are seen to be quite similar in appearance.

Figure 23 shows the correlation intensities obtained for OKLAHOMA SOONERS using a BPOF based on the letter O. The 4 occurrences of the letter O are clearly indicated by strong correlation peaks (remember that the image is inverted in the correlation plane). The values of autocorrelation intensities vary from 27.0 to 28.4, while the highest cross-correlation intensity is 5.64 (for the letter M). Correlations of the letter S with OKLAHOMA SOONERS are shown in Figure 24. Autocorrelation peak intensities are 26.6 and 28.9, while the highest cross-correlation peak is 7.92 (for the letter E), followed by 3.92 (for the letter O).

For the next simulation, the BPOF for the letter O was modified by blocking the central lobe, to achieve a high-pass effect. The resulting correlations of O with OKLAHOMA SOONERS is shown in Figure 25. The autocorrelation peak intensities range from 15.7 to 16.6, and the highest cross-correlation intensity is 2.32 (again the letter M, as for the full-bandwidth case). The high-pass filtering resulted in improving the discrimination by a factor of about 1.4 in this example. Since all the letters have about the same area, it is not expected that low frequency components are of much use in distinguishing one letter from another. Doing the same high-pass operation on the BPOF for the letter S produced the correlations shown in Figure 26. The two autocorrelation intensities were 13.35 and 14.90, while the highest cross-correlation intensity was 4.44 (for E, as before). The use of high-pass filtering results in a slight degradation of the discrimination ability in this instance. The average autocorrelation intensity to largest cross-correlation intensity reduces, by a factor of about 0.91.

The next operation investigated was low-pass filtering. In Figures 27 and 28, only the main lobe of the BPOFs were used. Figure 27 shows correlation intensities for the letter O with OKLAHOMA SOONERS, while Figure 28 is based on a BPOF central lobe for the letter S. The four O autocorrelation intensities range from 6.50 to 7.23, while the highest cross-correlation intensity is 3.09 (for the letter A), with others close behind. The two S autocorrelation intensities are 7.15 and 7.98, while the highest cross-correlation intensities are two occurrences of 3.37 (for letters E and O).

SLM HARDWARE AND SOFTWARE

The spatial light modulator (SLM) used as a filter was a liquid crystal television (LCTV), a Realistic model 16-156. This device's liquid crystal display consists of a matrix of 158 by 144 pixels on a 7.1 by 5.3 cm panel. The pixel spacing was approximately 445 by 365 μm . Two LCTVs of the same brand and model were purchased, and they were found to differ somewhat in detail. The data presented is for the one selected for use as a filter in the experimental work. Pixel spacing was approximately 445 by 365 μm . Approximately 80 percent of each pixel's area was active area, with the remaining 20 percent clear interpixel area in which optical beam phase rotations were not affected by voltages applied to the pixels.

As discussed in Tai [5], several modifications must be made to the LCTV to convert it into a useful SLM. A stop must be broken to enable the liquid crystal display to be positioned at a 90 degree angle to the main tv unit. A diffuser and two polarizers (of very low quality) must be removed from the display unit. What remains is a nematic liquid crystal sheet sandwiched between thin glass plates.

In order to use the LCTV as an SLM, it is also necessary to be able to use a computer to program the individual pixels. The LCTVs used were equipped with composite video input jacks. Available computers were DOS PCs. Interfacing circuitry is thus required to convert the video output of the computers into the format required by the LCTV video input. A circuit board, based on a Motorola MC1377 Color Television RGB to NTSC Encoder integrated circuit, was purchased and found to not work. After some design modification, a workable circuit was obtained to perform interfacing between the computer's digital RGB format and the LCTV's analog NTSC format. The interface circuitry allowed computer-generated images to be displayed on the LCTV screen, but addressing individual pixels required additional software. The CGA graphics screen consists of 200 by 640 pixels which address 133 by 149 (of 144 by 158) pixels of the LCTV. This mapping is due to the computer's video timing signal. In order to vertically map computer pixels to LCTV pixels, each two out of three computer rows were mapped to one row of LCTV pixels. There is no correspondingly good horizontal mapping, thus some horizontal blurring is unavoidable. The horizontal blurring

distortion was minimized by a trial-and-error process. The end result was that a fairly complex pattern of using particular computer pixels and skipping others was found to work quite well. Computer programs were written to allow desired patterns to be written on the LCTV.

LCTV CHARACTERIZATION AND PARAMETER OPTIMIZATION

Now that hardware and software was in place that permitted a desired pattern to be generated on a computer and transferred to the LCTV, the LCTV was tested extensively to determine optimum settings of the operational variables. These variables were the input polarization angle, output polarizer angle, and the amount of dc bias used with the liquid crystal panel. Several authors have given theory and data for inexpensive LCTVs [6,7,8,9]

The birefringent axes of the liquid crystal material rotate as the LCTV's dc bias voltage (television contrast control) is varied. To determine the orientation of birefringent axes, the bias voltage was varied from 13.4 to 19.4 volts dc in 0.2 volt steps. At each step the input beam polarization is varied until an analyzer at the output can obtain the most complete null possible. This is done both for pixels in the on state and pixels in the off state. The result of this measurement is shown in Figure 29. Measurements of a number of LCTV parameters versus contrast voltage are shown in Figure 30. Based on this data, the maximum difference in ON state and OFF state polarizations is 11° , occurring at a bias voltage of 18 V. The various angles and polarizations appropriate to this choice are shown in Figure 31. It is also important, however, to block the interpixel light, as this adds only noise at the correlation plane. Unfortunately, it is not possible to completely block interpixel light and still use the LCTV as a phase modulator. The best bias voltage for blocking interpixel light is 14.6 V. A vector diagram for this choice is shown in Figure 32. It is seen that an ON-OFF phase difference of only 2° is obtained at this bias point. Since it is not possible to simultaneously maximize the ON-OFF phase difference and eliminate interpixel light, a compromise bias voltage of 16.4 V was used for the experimental work. As shown in Figure 33, this choice produces an ON-OFF phase difference of 7° , but unfortunately passes 82 percent of the interpixel light. The net result is that correlation plane images were superpositions of the original image and correlation intensities.

One concern usually voiced for using inexpensive LCTVs as SLMs is that the phase is not uniform over the extent of the display. This difficulty has been overcome by mounting the LCTV in a liquid gate [10], sandwiching the device between optical plates [11], correcting the phase holographically [12], and encoding a correction into a BPOF [13]. Horner [14] suggests that phase correction might not be necessary if phase distortion is on the order of 1.6λ or less. Interferometric measurements on the LCTV used in this experimental system indicated that its phase

Distortion was not much worse than this, so no phase correction was used in the experimental work described in this report.

CORRELATOR TEST BED DESIGN

The basic configuration to be used is the standard 4f or Vanderlugt correlator. The minimum focal length that can be used is dictated by the pixel sizes and overall extents of the input and filter patterns. Using the rather large pixel spacing of the liquid crystal tv (LCTV), 365 x 445 μm , and an overall extent of 120 x 146 pixels for both input and filter, a minimum focal length to avoid aliasing is calculated to be approximately 20 m. Design constraints were that the system should use only available 2 inch diameter lenses, all tolerances should be correctable by positioning adjustments, and the overall length should be less than 8 m (to fit on an laboratory countertop). The desired 20 m focal length was achieved in a shorter distance by using lens pairs in a telephoto configuration, and the overall length was further reduced by using a phase correction lens at the filter plane. The final correlator design is shown in Figure 34. All lenses are mounted on rails, permitting all fine-tuning adjustments to be made by adjusting lens positions. Because of the LCTV's large pixel size, the overall length is somewhat longer than usually used in correlator test beds, but the design proved to be quite workable.

EXPERIMENTAL CORRELATOR RESULTS

Although input transparencies were prepared for additional images, the only image used in experimental work was that of a simple rectangle, a slit 6.935 mm tall and 1.422 mm wide. The rectangular aperture was constructed using four razor blades. Because of delays in getting the system operational, no attempt was made to take data for more complex inputs (some parts ordered for use in the correlator system did not arrive until after all work had been completed). In order to ensure that the image's Fourier transform scale matched that of the filter written on the LCTV, the calculated Fourier transform was drafted on paper (eminently doable, thanks to the relatively large scale involved), and the size of the rectangle was trimmed to make the transform-plane pattern agree with the calculated pattern. All final adjustments, such as small adjustments in the x-y position of the LCTV, were made while observing the correlation plane output, captured by a frame grabber from the CCD camera and displayed on a computer monitor.

The image observed at the correlation plane had a noticeable interference bar pattern, with a slight vibration effect also noticeable. The very small pixel size of the CCD camera thus resulted in a correlation peak intensity that varied with time. The relative time during a LCTV scan cycle at which an image was captured may also have been a consideration. Figures 35, 36, and 37 show three examples of detected correlation plane intensities.

Some averaging of adjacent pixels had to be done to accommodate restrictions imposed by the 3-dimensional mesh plotting program, but the overall appearance is little changed. Figure 35 shows a correlation peak of 220, with a background average noise level of 45.35. The SNR is thus 4.85. The average of intensities below 50 percent of the peak is 51.08, thus the PSR is 4.3 for the image of Figure 35. For Figure 36, the correlation peak intensity was 255, the noise level was 47.45, and the average of intensities below 50 percent of the peak was 51.11. The resulting ratios are then PSR = 4.99, SNR = 5.37. For Figure 37, the correlation peak intensity was 234, the noise level was 47.54, and the average of intensities below 50 percent of the peak was 51.03. Ratios are PSR = 4.59, SNR = 4.92. The appearances of the correlation plane images differ considerably from the ideal of sharp narrow correlation peaks, but at least a correlation peak is clearly evident. The image of the input rectangle, transmitted as interpixel light, is also evident. Probable factors leading to a reduced correlation peak include decorrelation at higher spatial frequencies due to a slight scale mismatch, defocussing due to nonuniform phase distribution across the LCTV aperture, scattering due to use of a low-quality polaroid-type polarizer, and perhaps some system misalignment due to vibration, as the 8 m long system was not installed on an anti-vibration table.

High-pass and low-pass operations were performed by physically blocking parts of the filter aperture. Figure 38 is an example of the effect of high-pass filtering. The zero order main lobe of the filter response was blocked for this measurement. The result was a correlation peak of 86, separating two other peaks. Noise background intensity was 47.8, the SNR is 1.8. There are essentially no intensities below 50 percent of the peak, so no PSR can be calculated. Figure 39 shows the results of a low-pass experiment. A variable circular aperture was placed after the LCTV to achieve a low-pass filtering effect. This particular experiment had a correlation intensity peak of 165, with a background noise level of 48.11. The average of intensities below 50 percent of the peak was 50.09. The resulting PSR is 3.29, and the SNR is 3.43.

Lastly, Figure 40 shows a cross-correlation with another unspecified rectangle. This graph shows two peaks of not too different height. The taller has an intensity of 144. The background noise intensity is 46.82. One interesting feature that shows up well in this graph is the slit image, caused by interpixel light transmission.

CONCLUSIONS AND RECOMMENDATIONS

The simulations that have been done suggest that low-pass filtering could play a useful role in applications such as target recognition. The low-pass correlation produces an indication that a blob of about the right size is present, and could be followed by a more detailed examination if suitable candidates are indicated.

High-pass filtering produces a sharper correlation peak, but at the expense of heightened sensitivity to scale changes and rotations. A high-pass filtered correlation is likely to find use only if a very specific size and orientation are being sought. Bandpass filtering is unlikely to be of much use unless the major differences between similar image objects lies in a narrow range of spatial frequencies.

The large pixel size, lack of resolution, and small dynamic range all conspire to restrict the LCTV to demonstration-type uses. The fact that interpixel light could not be eliminated in the unit employed was a major disappointment. The fact that LCTV pixels are rectangular, not square, is an annoyance. Still, with further development, it is clear that the LCTV could be a useful component in a correlator test bed. One further development that would be interesting to pursue would be to develop a program to cause the LCTV-written filter to optimize itself.

ACKNOWLEDGEMENTS

Graduate student Scott Coffin did most of the experimental work, and his MS thesis was based on work supported by this project. Graduate student Chong Lee helped with part of the characterization of the LCTV, the development of the RGB to NTSC conversion electronics, and the preparation of input photographic transparencies. The support of the U. S. Air Force through Universal Energy Systems is gratefully acknowledged.

REFERENCES

1. Horner, J. L., and P. D. Gianino, Phase-only matched filtering, *Applied Optics* 23, 812-816, 15 Mar 1984
2. Gianino, P. D., and J. L. Horner, Additional properties of the phase-only correlation filter, *Optical Engineering* 23, 695-697, Nov-Dec 1984
3. Horner, J. L., and J. R. Leger, Pattern recognition with binary phase-only filters, *Applied Optics* 24, 609-611, 1 Mar 1985
4. Oppenheim, A. V., and J. S. Lim, The importance of phase in signals, *Proc. IEEE* 69, 529-541, May 1981
5. Tai, A. M., Low-cost LCD spatial light modulator with high optical quality, *Applied Optics* 25, 1380-1382, 1 May 1986
6. Liu, H-K, J. A. Davis, and R. A. Lilly, Optical-data-processing properties of a liquid-crystal television spatial light modulator, *Optics Letters* 10, 635-637, December 1985
7. Bates, B., P. C. Miller, and W. Luchuan, Liquid crystal TVs in speckle metrology: optimum conditions for bipolar phase modulation, *Applied Optics* 28, 1969-1971, 1 June 1989
8. Liu, H-K, and T-H Chao, Liquid crystal television spatial light modulators, *Applied Optics* 28, 4772-4780, 15 Nov 1989
9. Lu, K., and B. E. A. Saleh, Theory and design of the liquid crystal TV as an optical spatial phase modulator, *Optical Engineering* 29, 240-246, March 1990
10. Bates, B., and P. C. Miller, Liquid crystal television in speckle metrology, *Applied Optics* 27, 2816-2817, 15 July 1988
11. Hughes, K. D., S. K. Rogers, J. P. Mills, and M. Kabrisky, Optical preprocessing using liquid crystal televisions, *Applied Optics* 26, 1042-1044, 15 Mar 1987
12. Casasent, D., and S-F Xia, Phase correction of light modulators, *Optics Letters* 11, 398-400, Jun 1986
13. Davis, J. A., D. M. Cottrell, J. E. Davis, and R. A. Lilly, Fresnel lens-encoded binary phase-only filters for optical pattern recognition, *Optics Letters* 14, 659-661, 1 Jul 1989
14. Horner, J. L., Is phase correction required in SLM-based optical correlators?, *Applied Optics* 27, 436-438

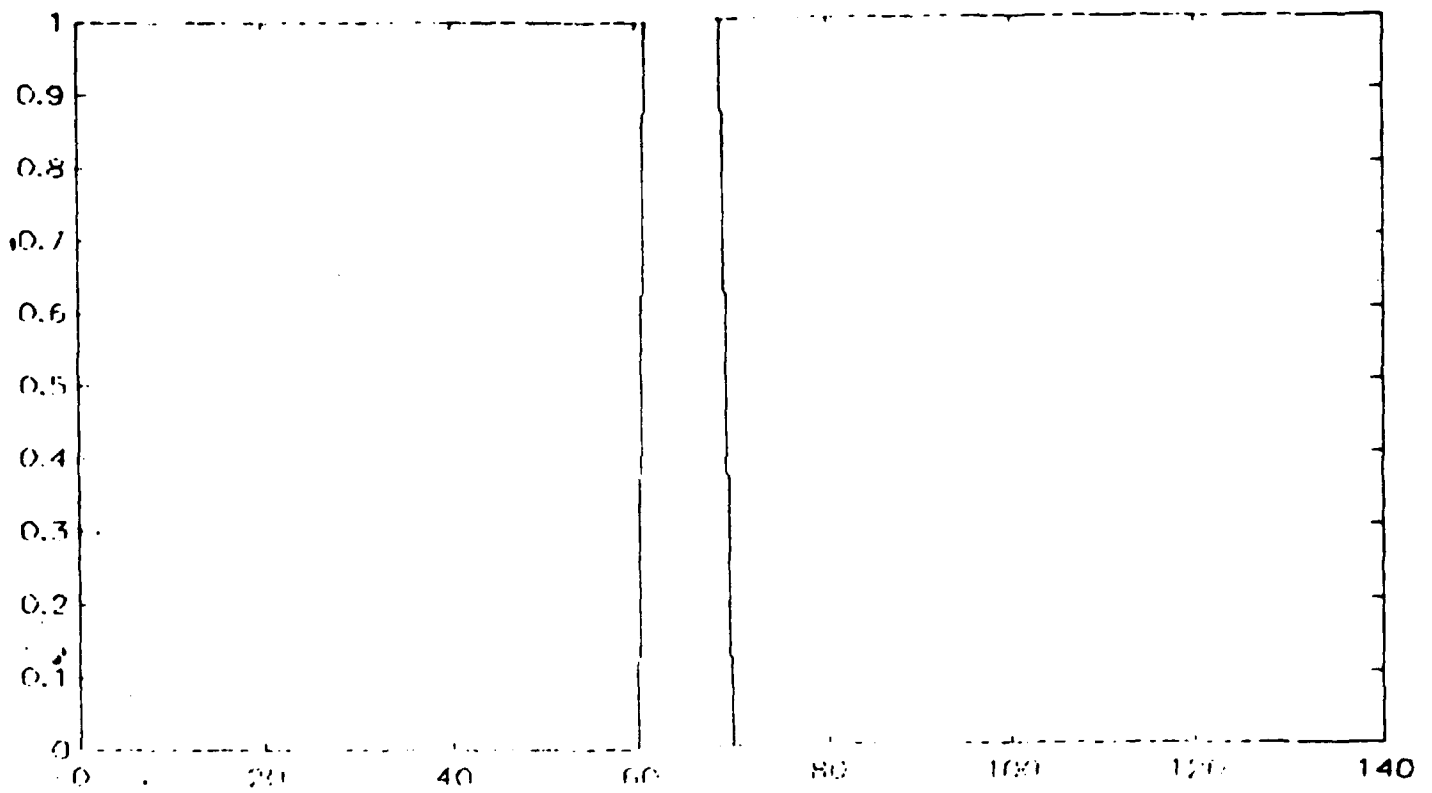


Figure 1. A 1-dimensional rectangular pulse waveform used to demonstrate the basic idea of a binary phase-only filter.

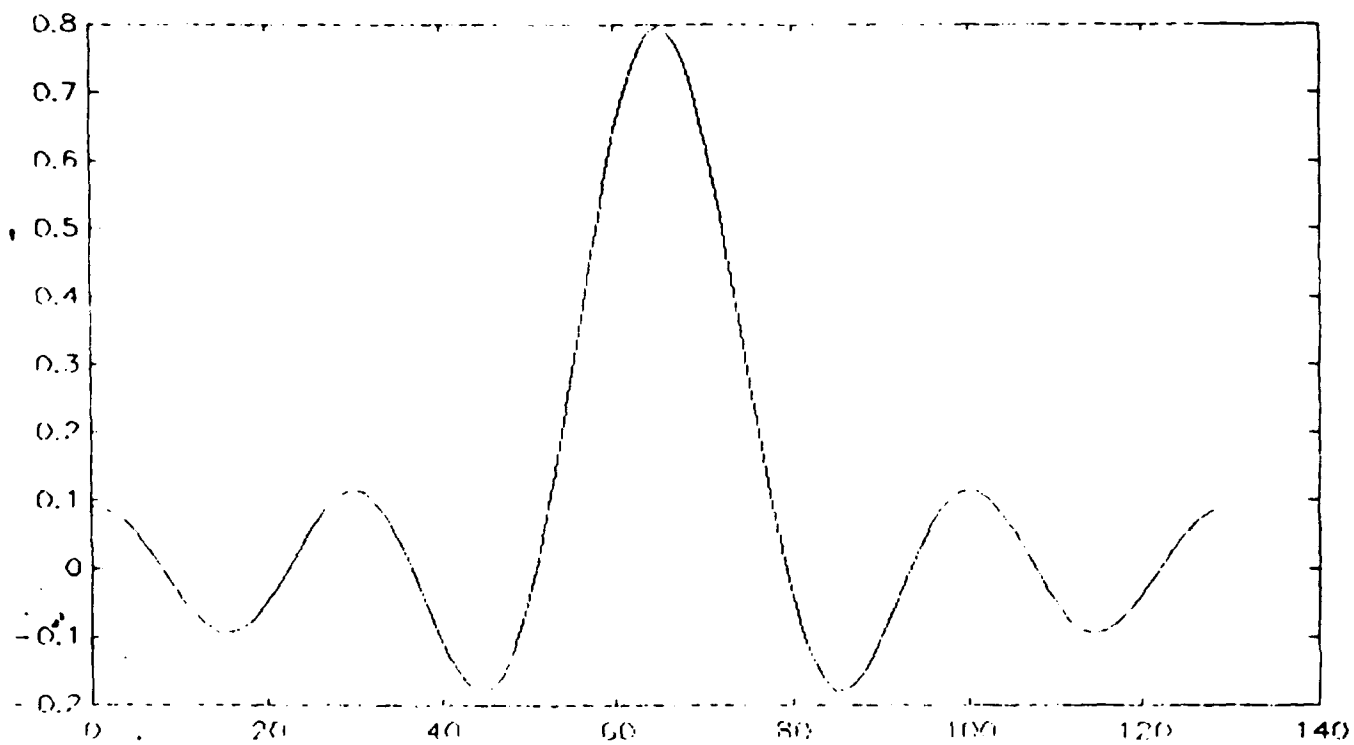


Figure 2. The Fourier transform of the pulse waveform of Figure 1.

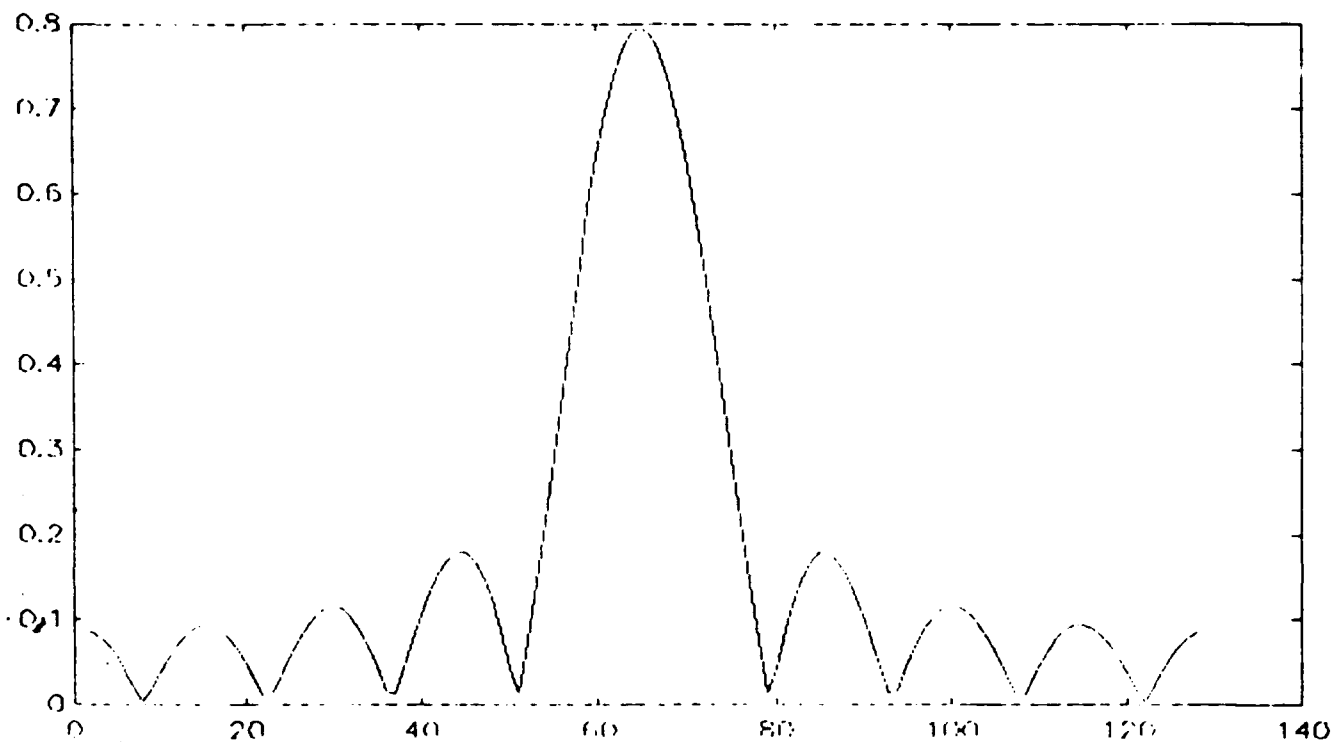


Figure 3. The magnitude of the Fourier transform of the pulse waveform of Figure 1 (result of a binary phase-only operation).

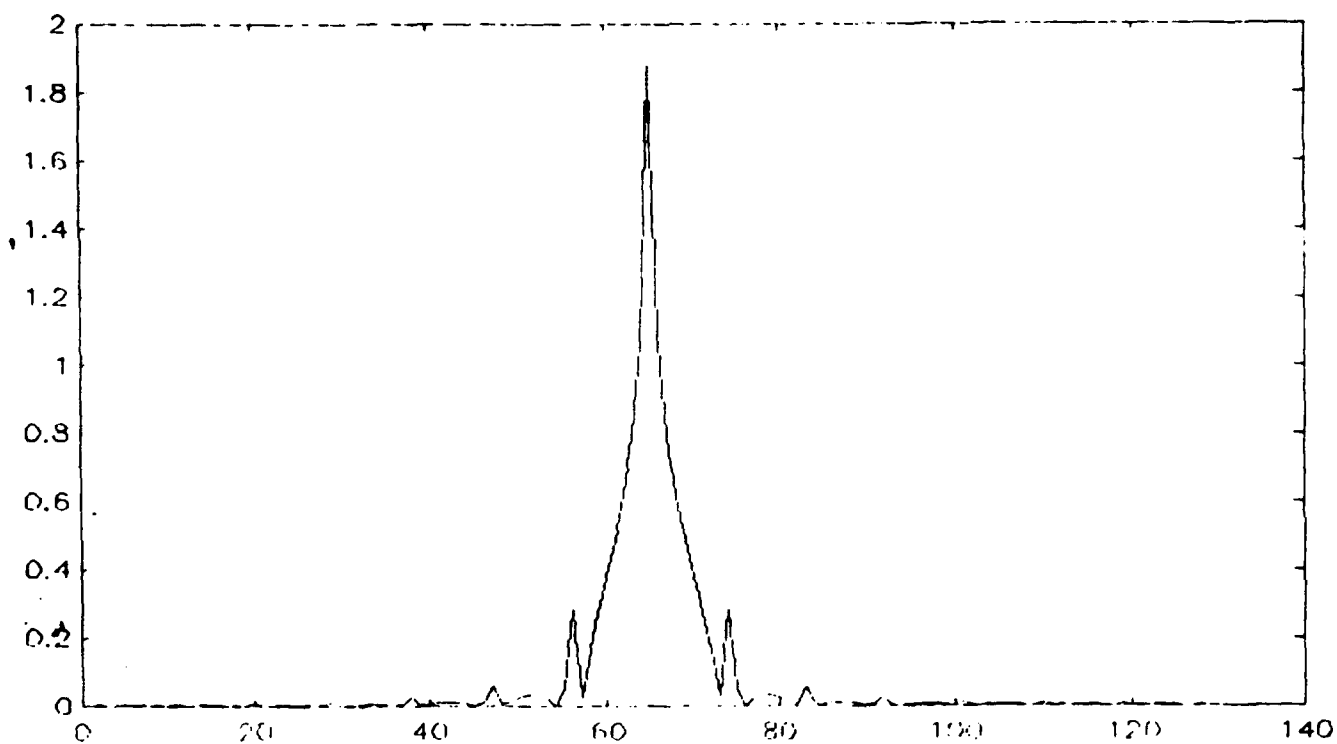


Figure 4. Fourier transform of the magnitude of the transform of the pulse of Figure 1, equivalent to a correlation using a binary phase-only filter.

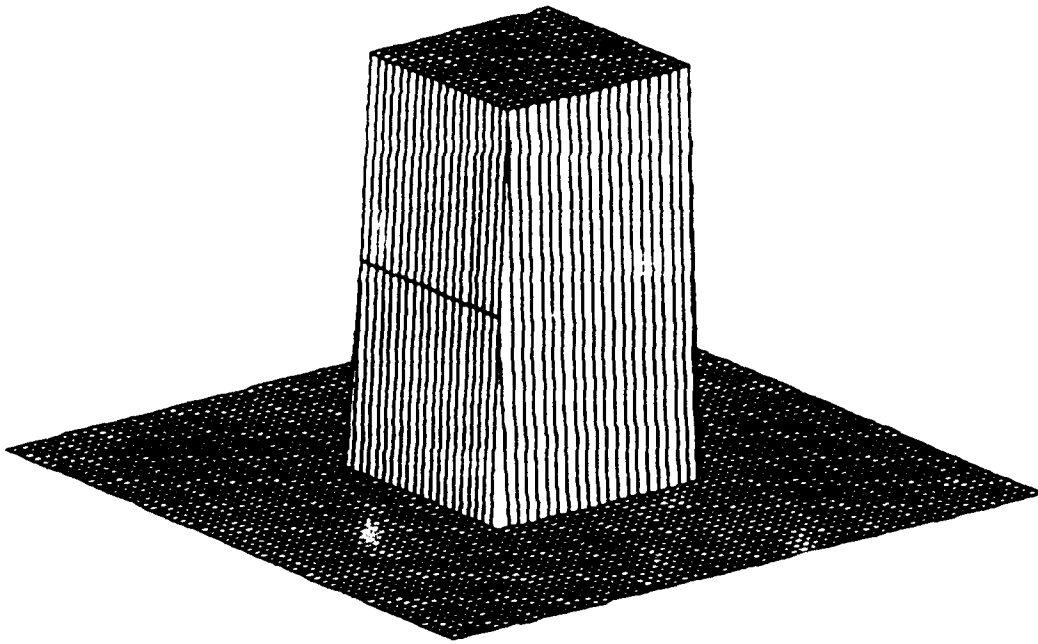
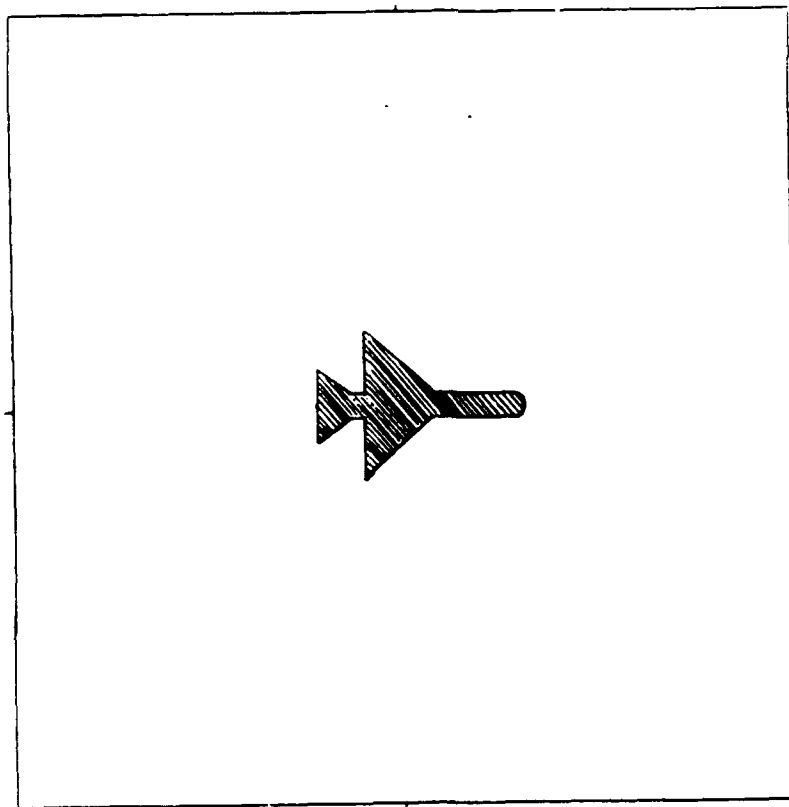


Figure 5. Relief graph showing the square pattern used in simulations. The square is 39x39 pixels in a 128x128 array.



5-17

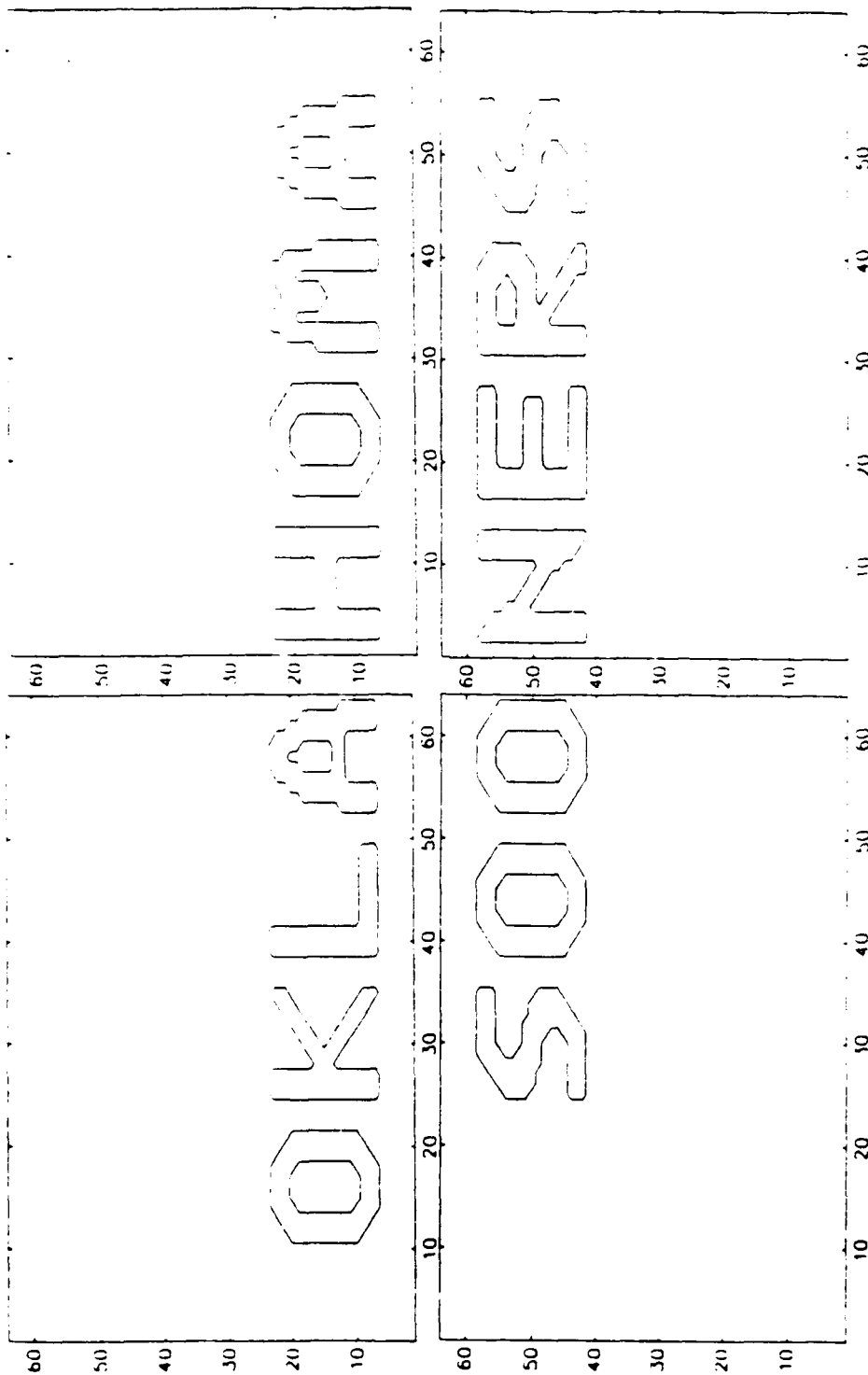


Figure 7. Contour graph of the text pattern used in simulations.

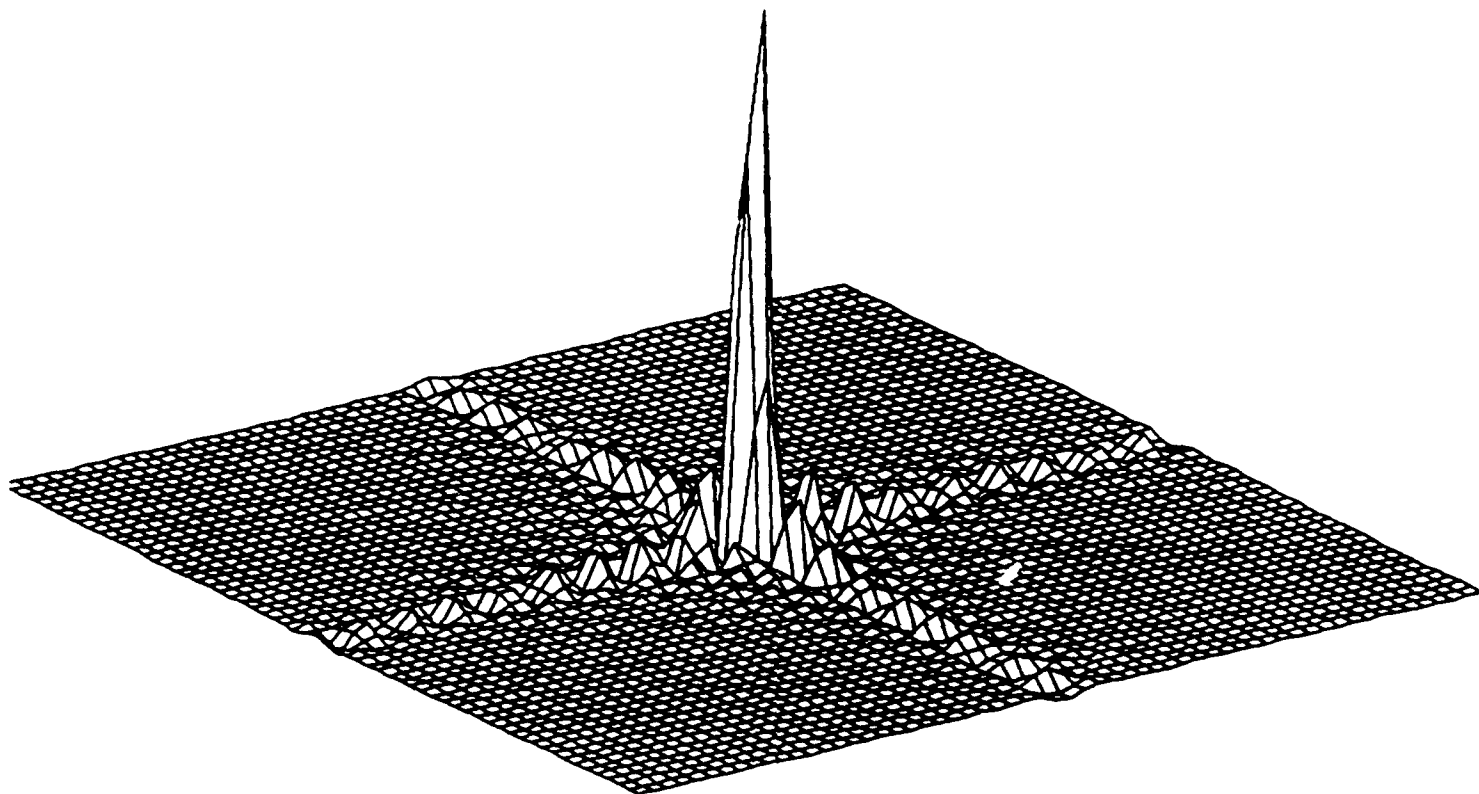


Figure 8. Fourier transform magnitude for the square shown in Figure 5.

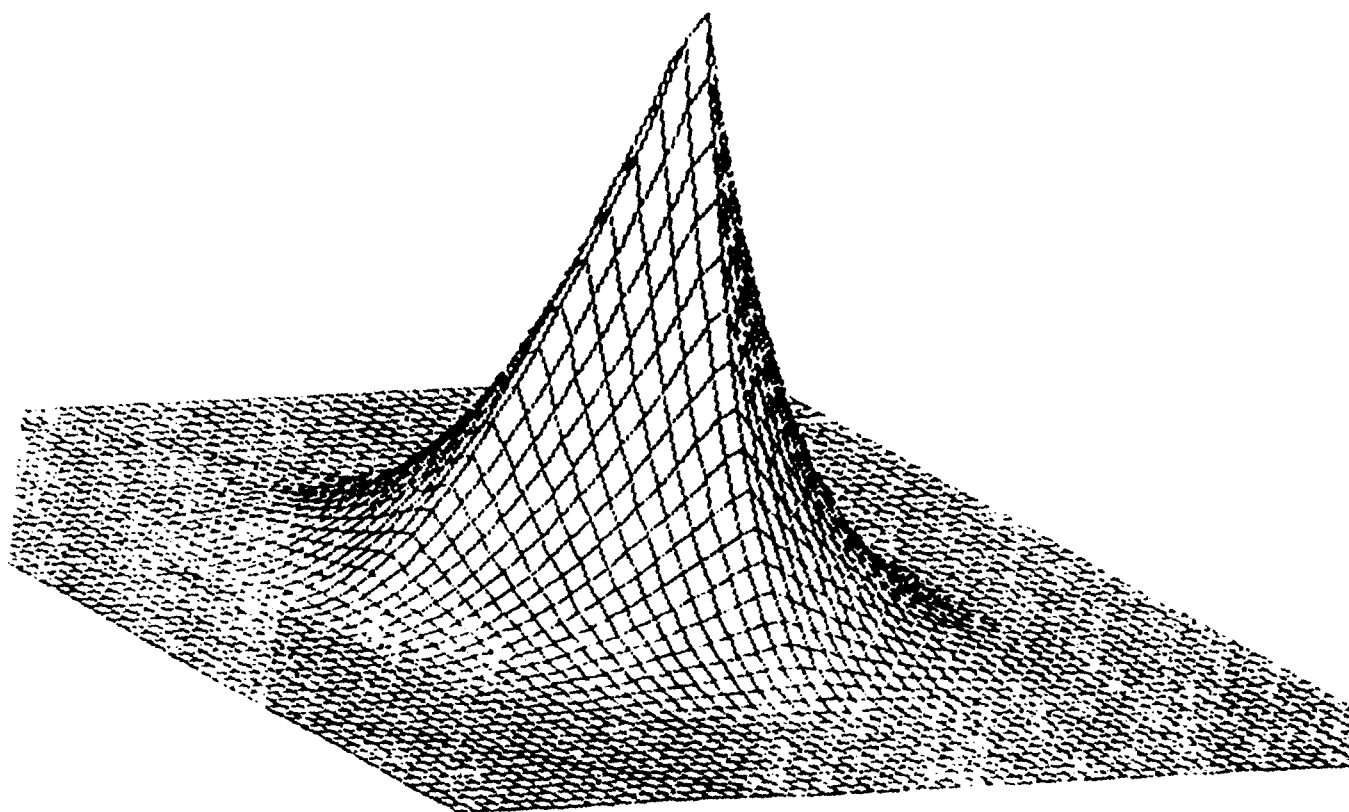


Figure 9. Correlation of the rectangular pattern of Figure 5 with itself using a matched filter. The correlation peak intensity is 1.00.

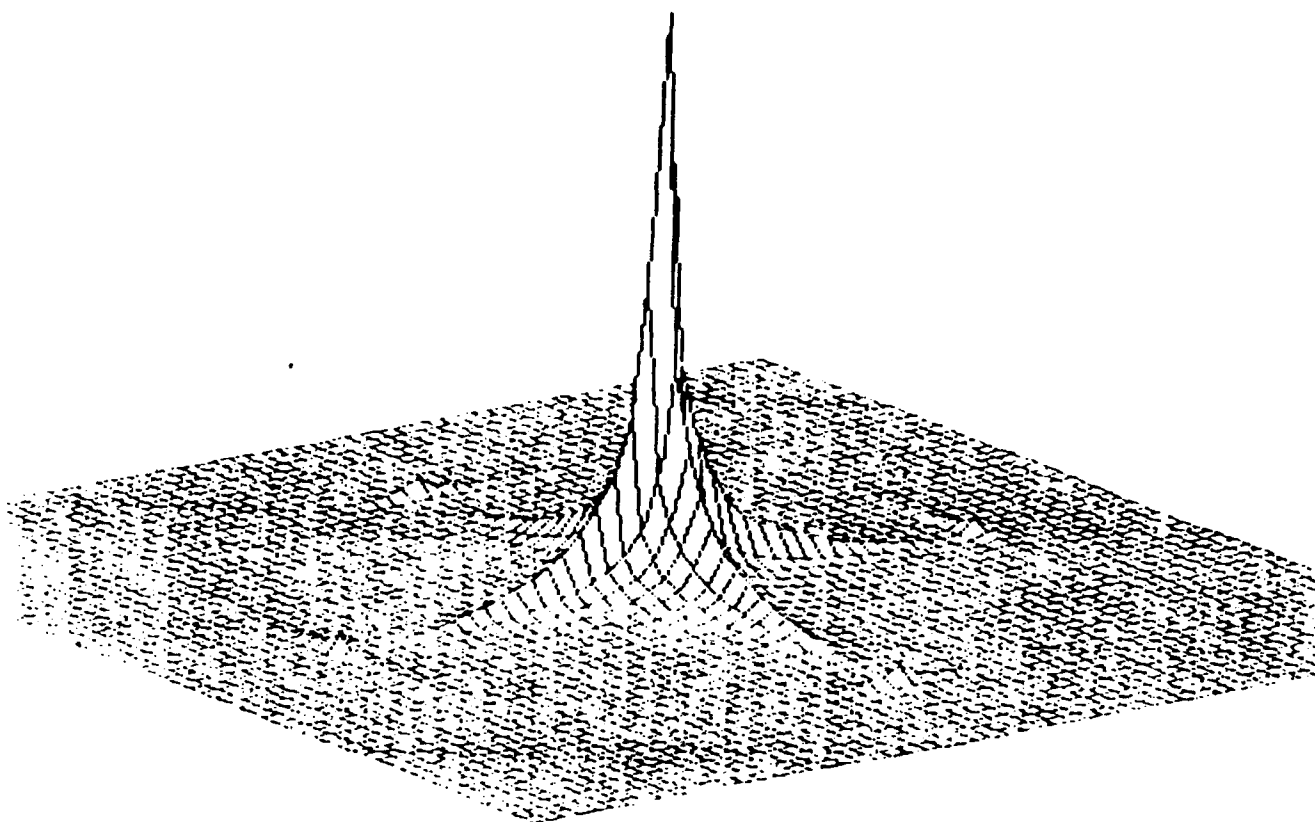


Figure 10. Correlation of the rectangular pattern of Figure 5 with itself using a binary phase-only filter. The correlation peak intensity is 38.09.

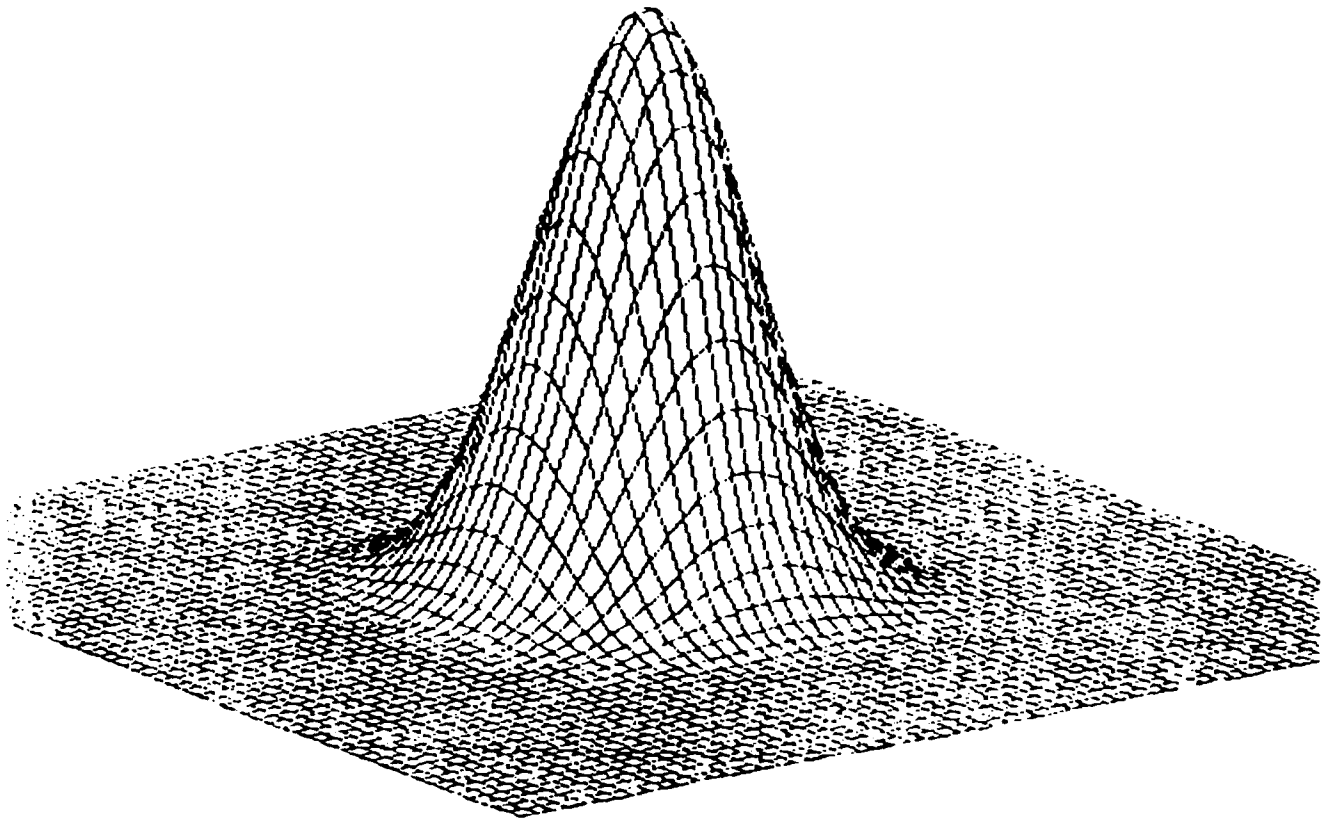


Figure 11. Correlation of the rectangular pattern of Figure 5 with itself using a low-pass processed BPOF. Only the main lobe of the Fourier response is kept. The correlation intensity peak is 1.95.

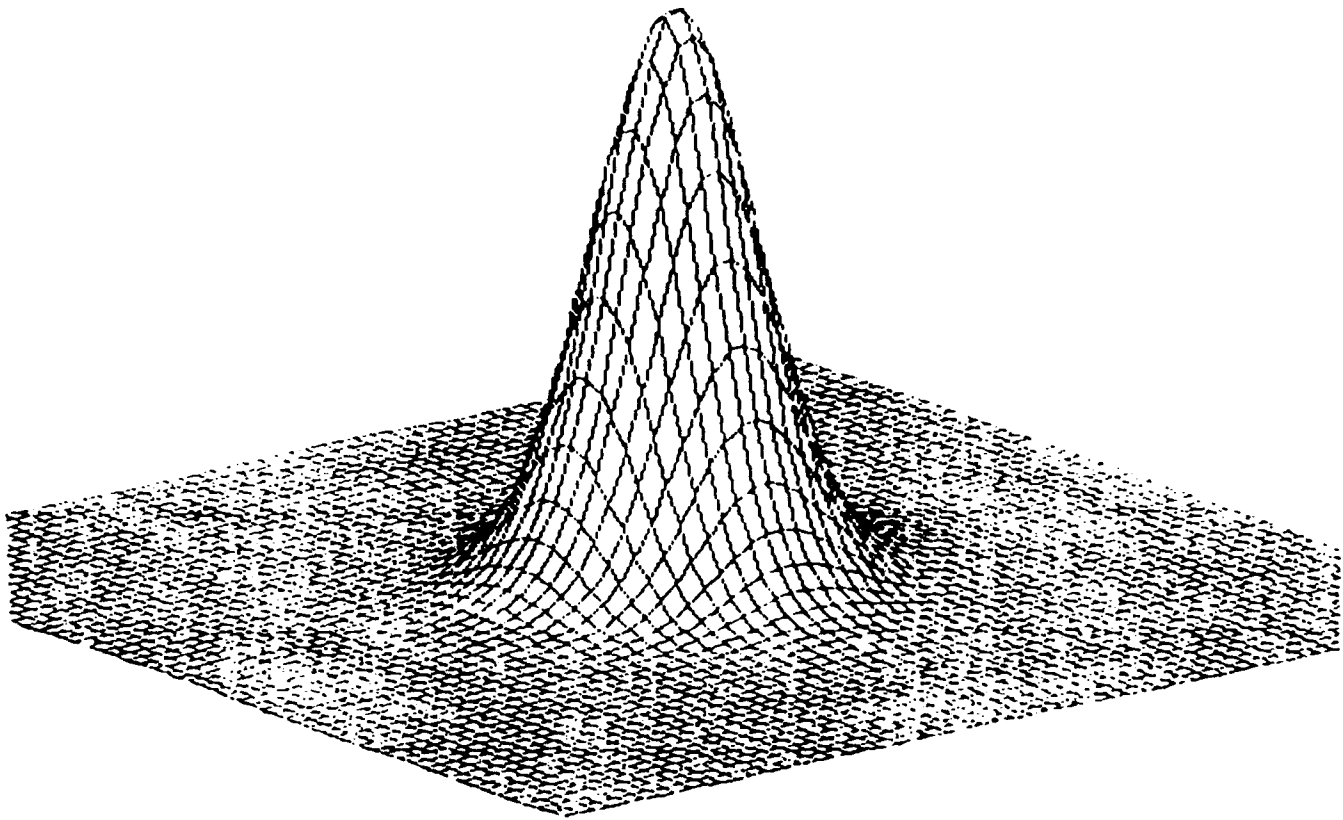


Figure 12. Correlation of the rectangular pattern of Figure 5 with itself using a low-pass filtering operation with a BPOF. Only the main lobe and the next lobes out are retained. The correlation intensity peak value is 2.71.

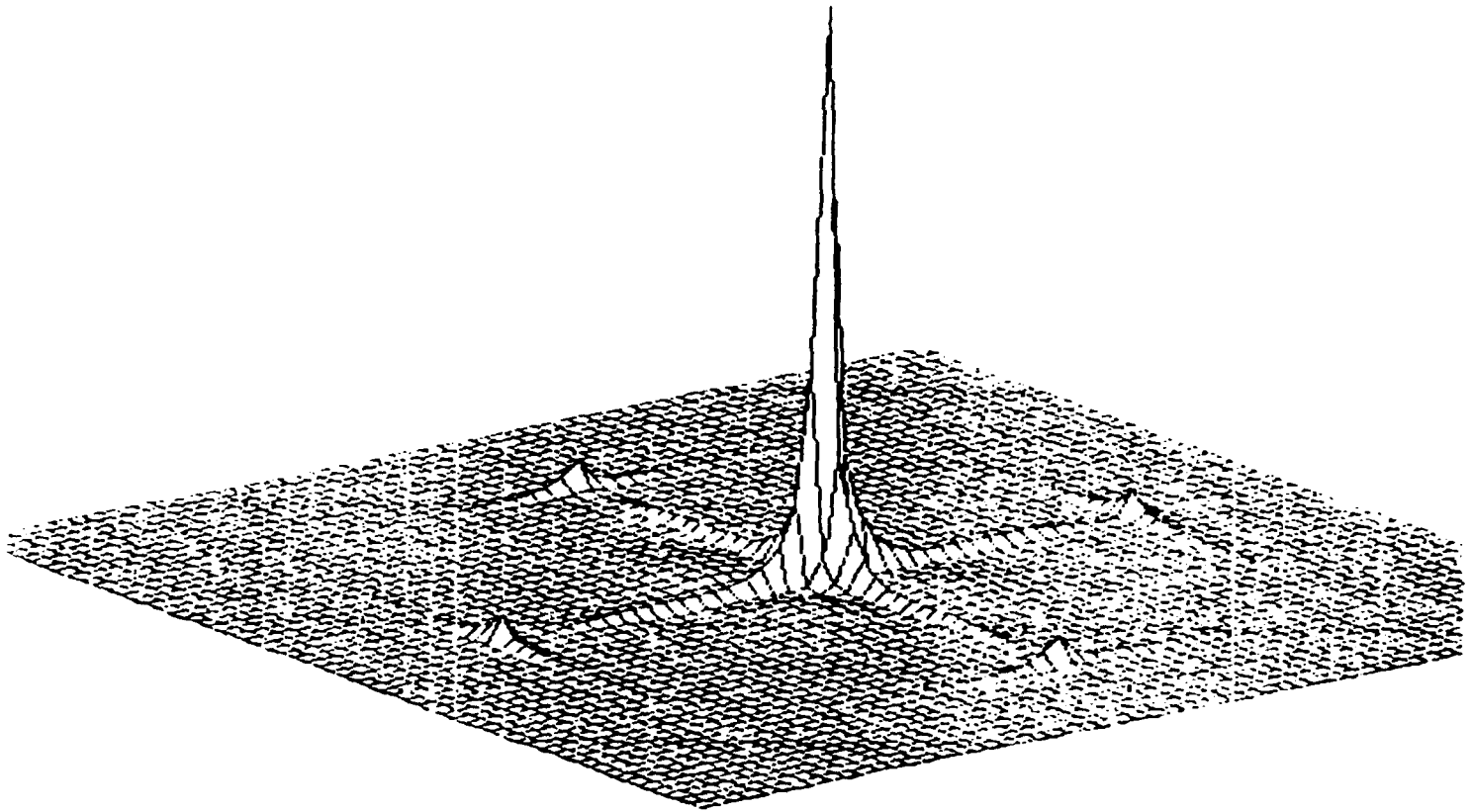


Figure 13. Correlation of the rectangular pattern of Figure 5 with itself using a high-pass filtering operation with a BPOF. Only the central lobe of the Fourier response is suppressed. The correlation intensity peak is 22.8.

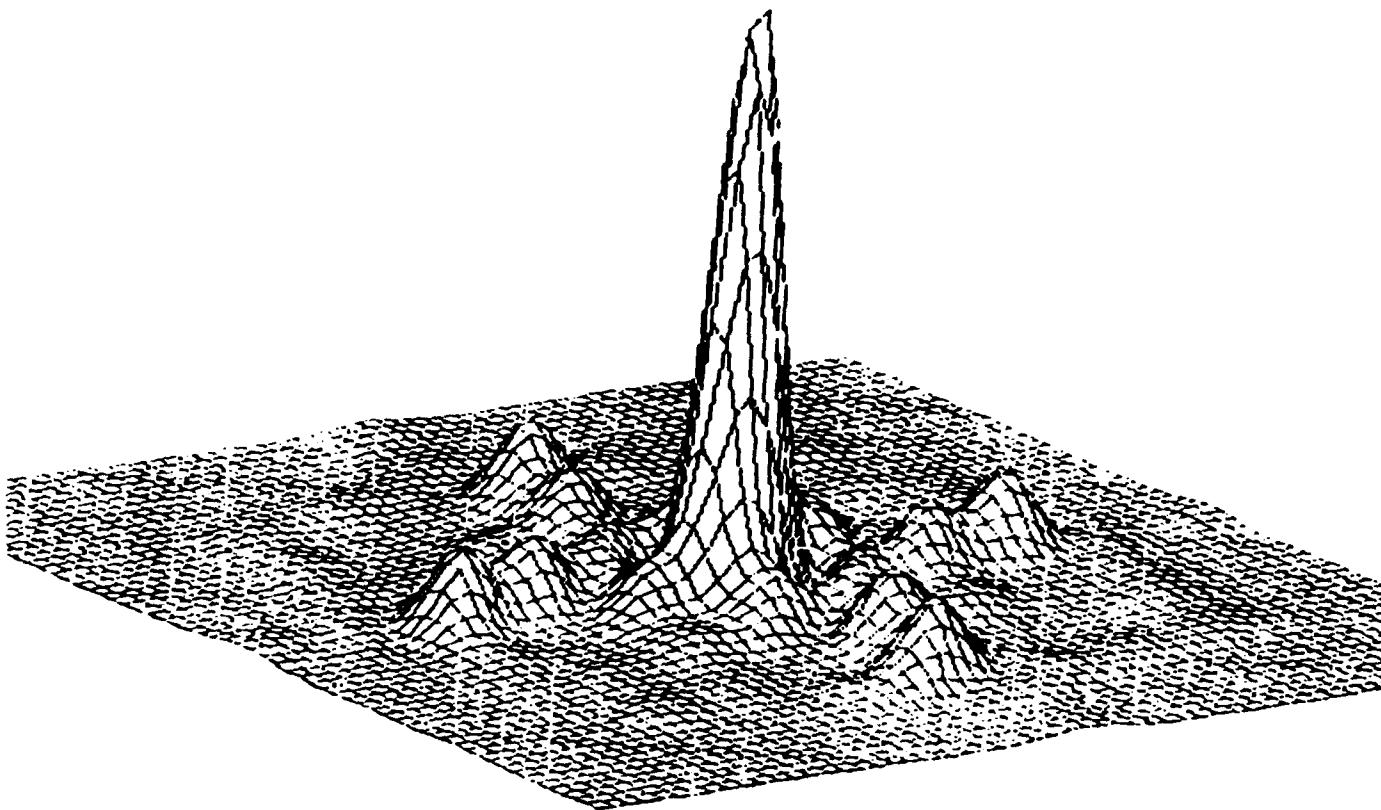


Figure 14. Correlation of the rectangular pattern of Figure 5 with itself using a bandpass filtering operation with a BPOF. Only the 2nd, 3rd, and 4th lobes of the Fourier transform response are used. The correlation intensity peak is only 0.26.

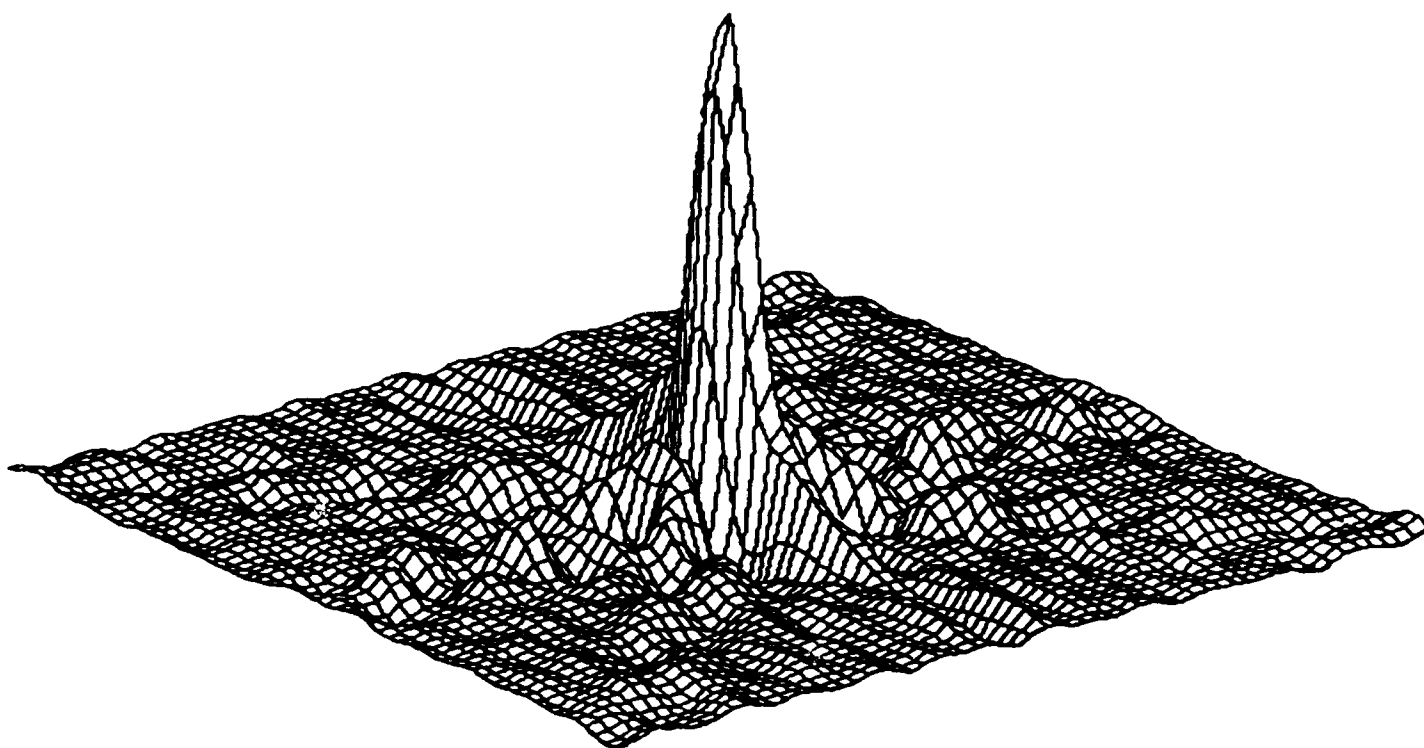


Figure 15. Magnitude of Fourier transform for airplane image of Figure 6.

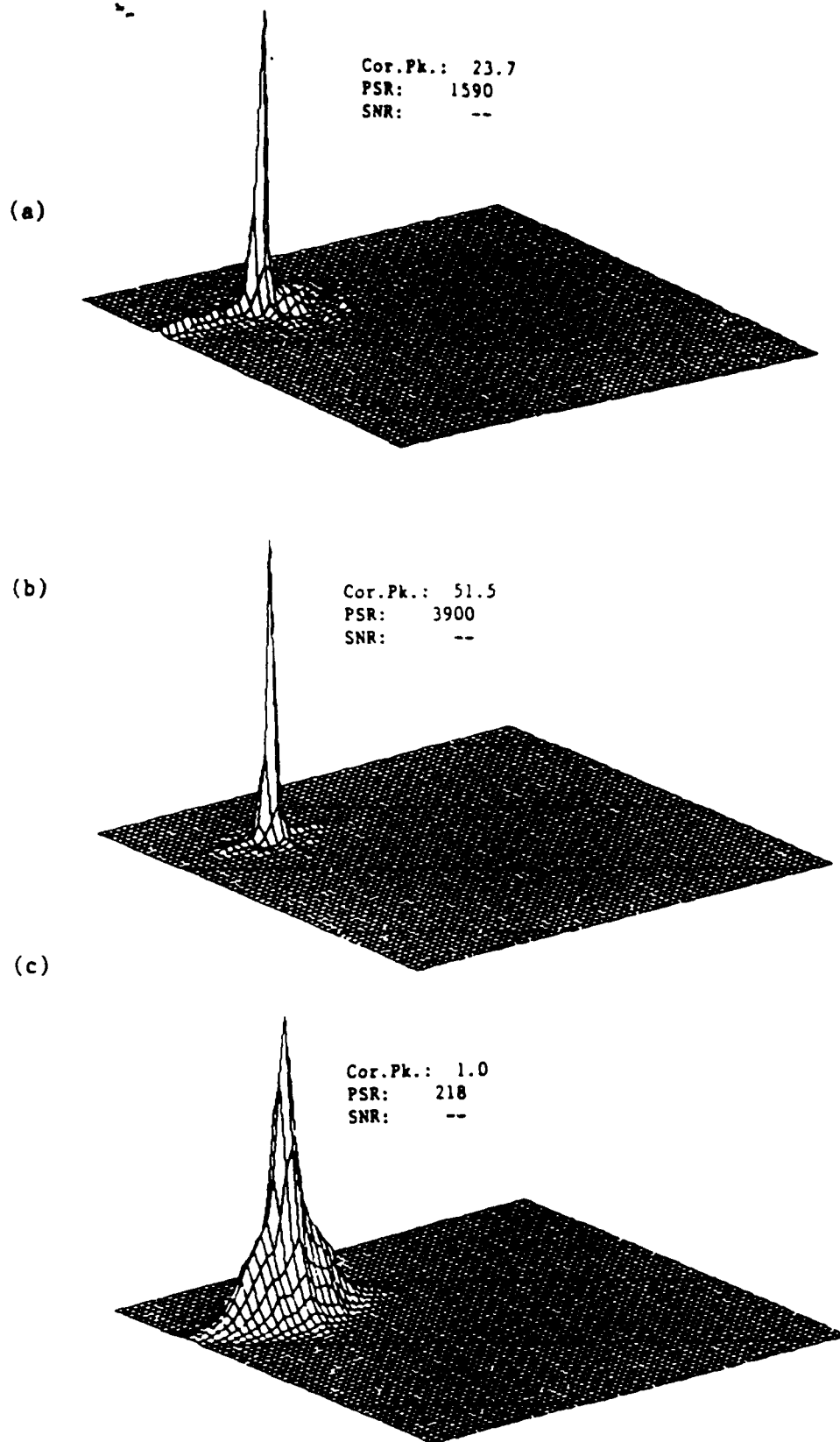


Figure 16. Correlations of an offset plane image with an (a) BPOF, (b) CPOF, and (c) MF for a centered plane image. Each graph is normalized to the same amplitude.

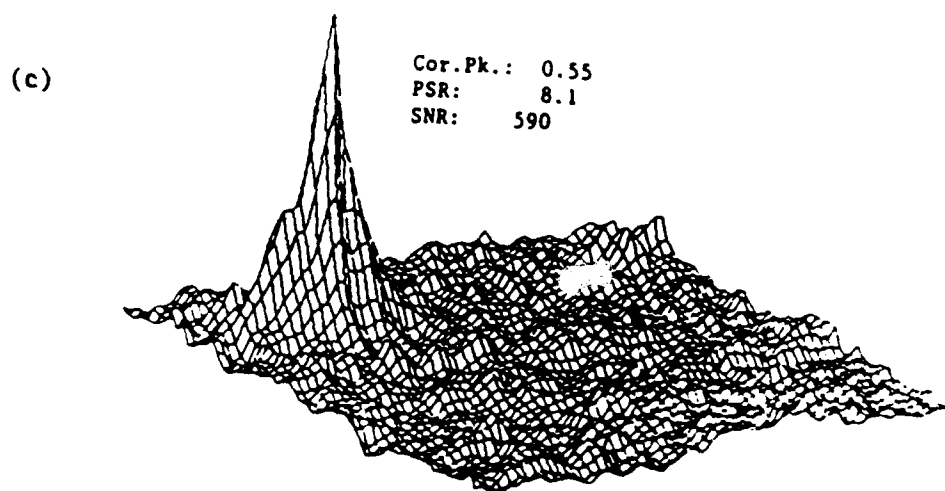
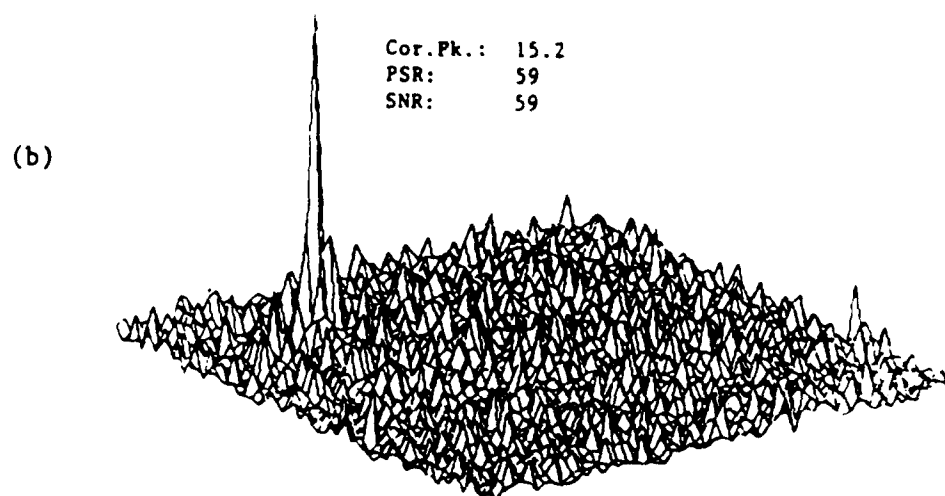
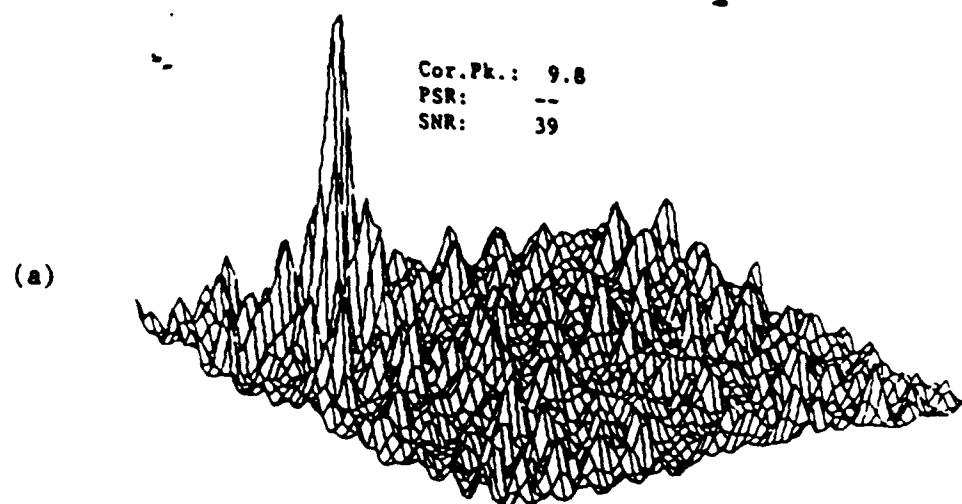


Figure 17. Same as Figure 16, but with added noise.

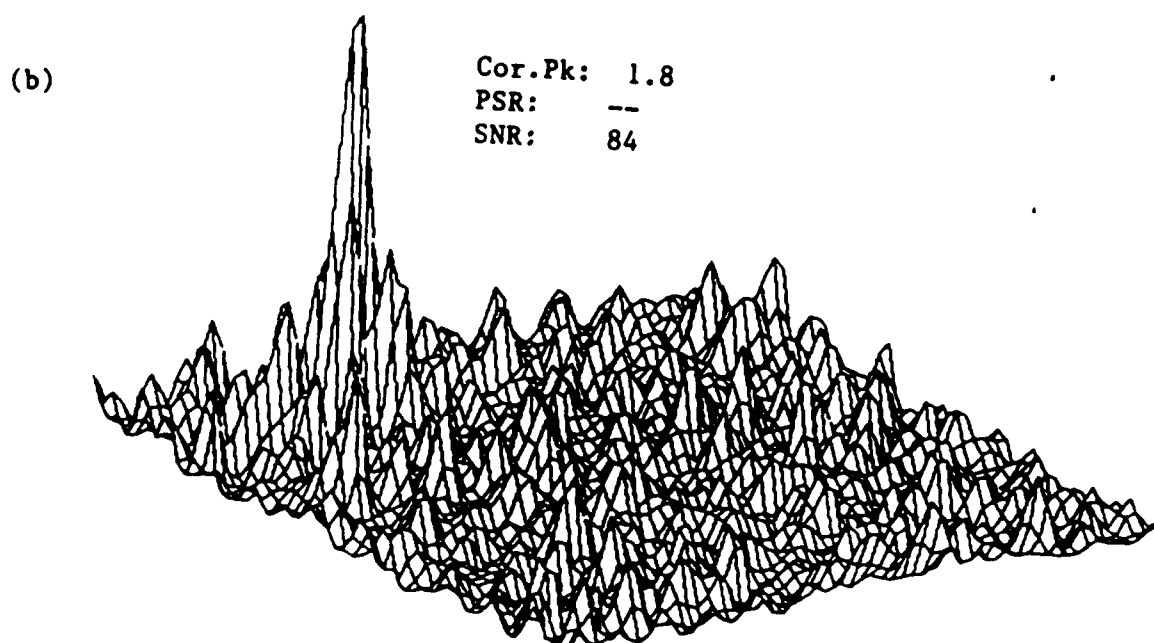
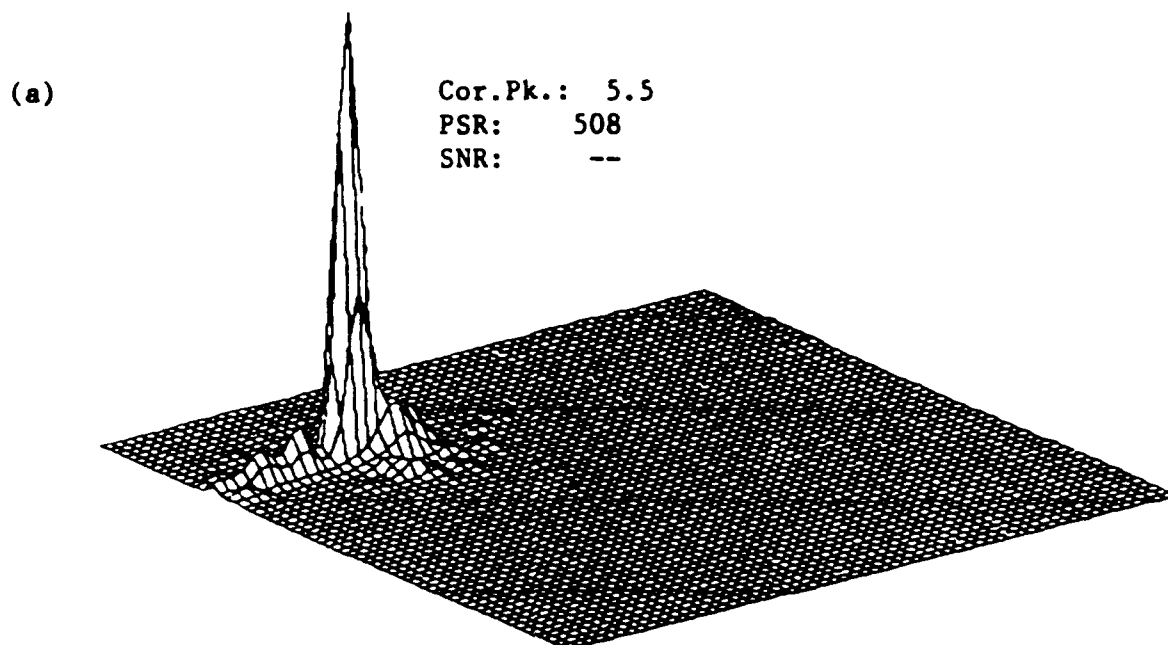


Figure 18. Correlation of an offset plane image with a reduced-bandwidth BPOF for (a) noise-free case, and (b) added noise case.

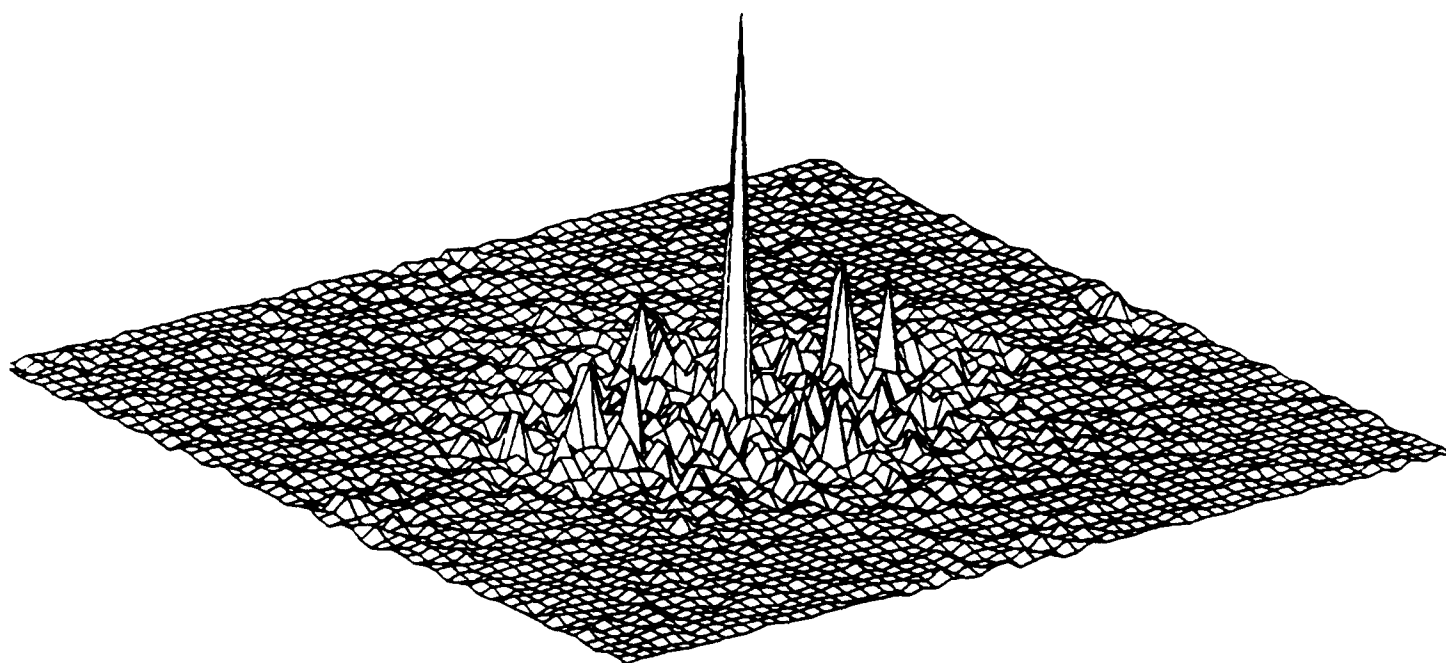


Figure 19. Fourier transform magnitude for OKLAHOMA SOONERS text pattern.

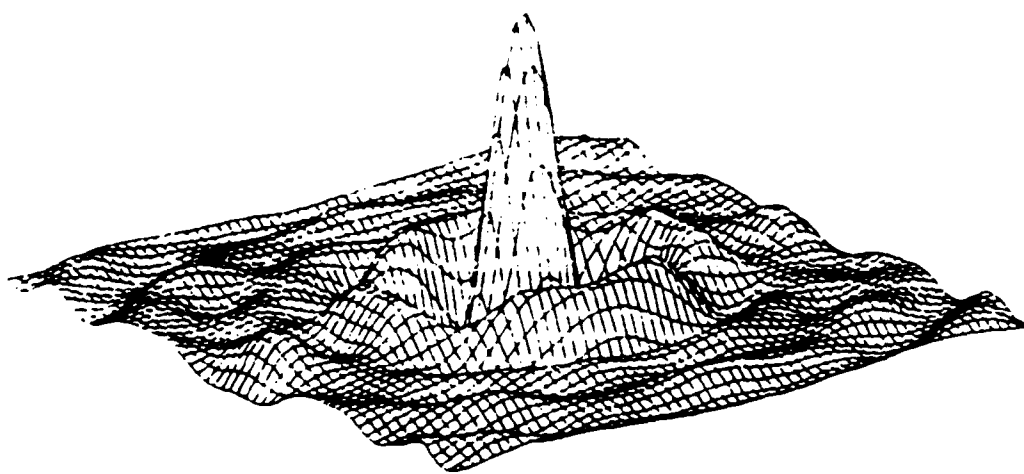
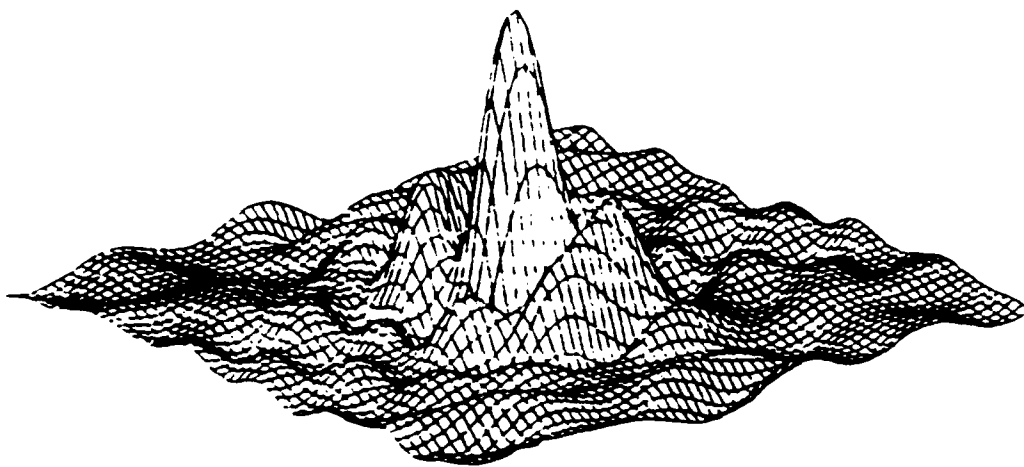


Figure 20. Fourier transform magnitudes for OKLAHOMA SOONERS letter "S" (top) and "O" (bottom).

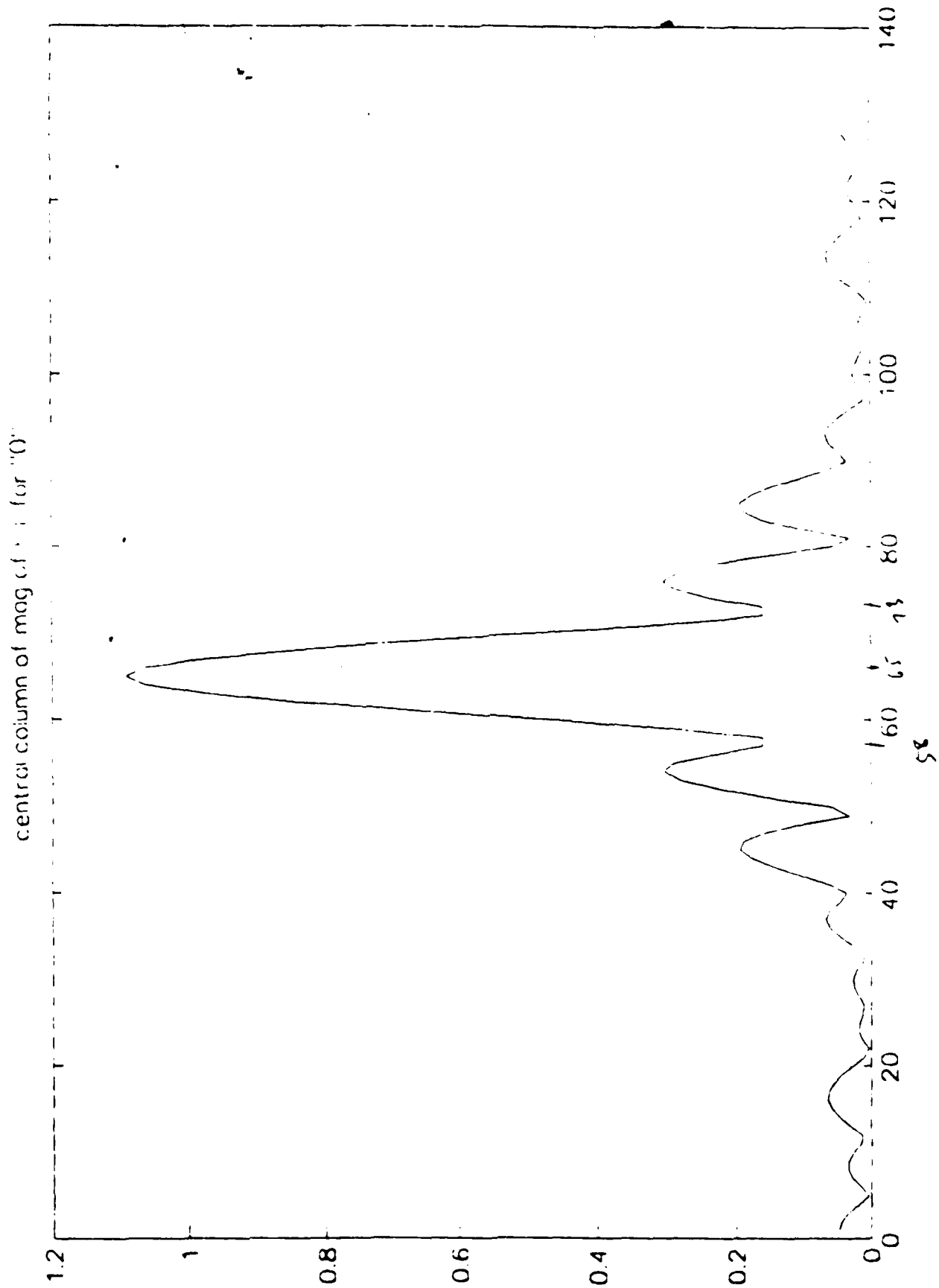


Figure 21. Graph of the central column of the Fourier transform magnitude for letter "O".

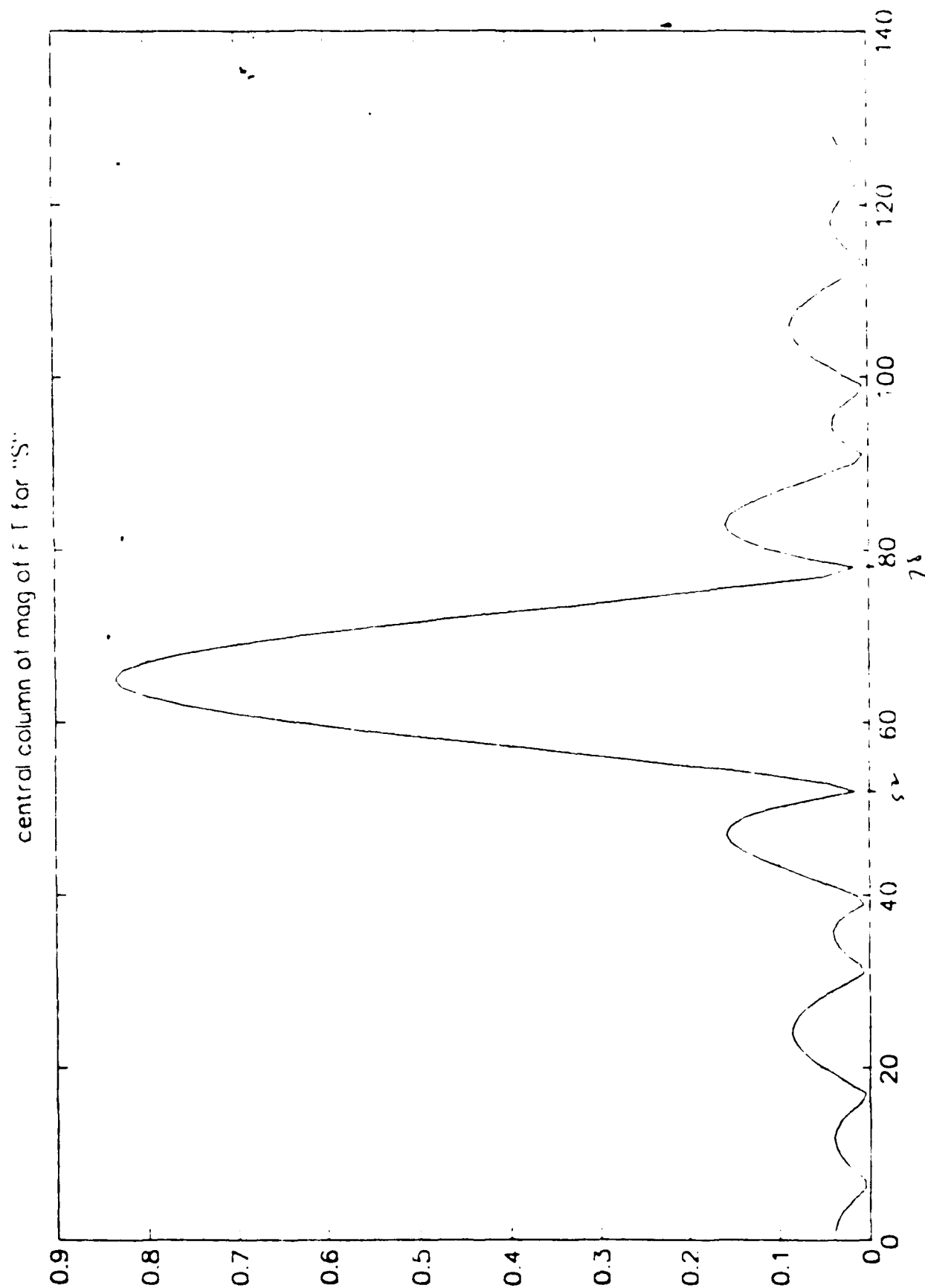


Figure 22. Graph of the central column of the Fourier transform magnitude for letter "S".

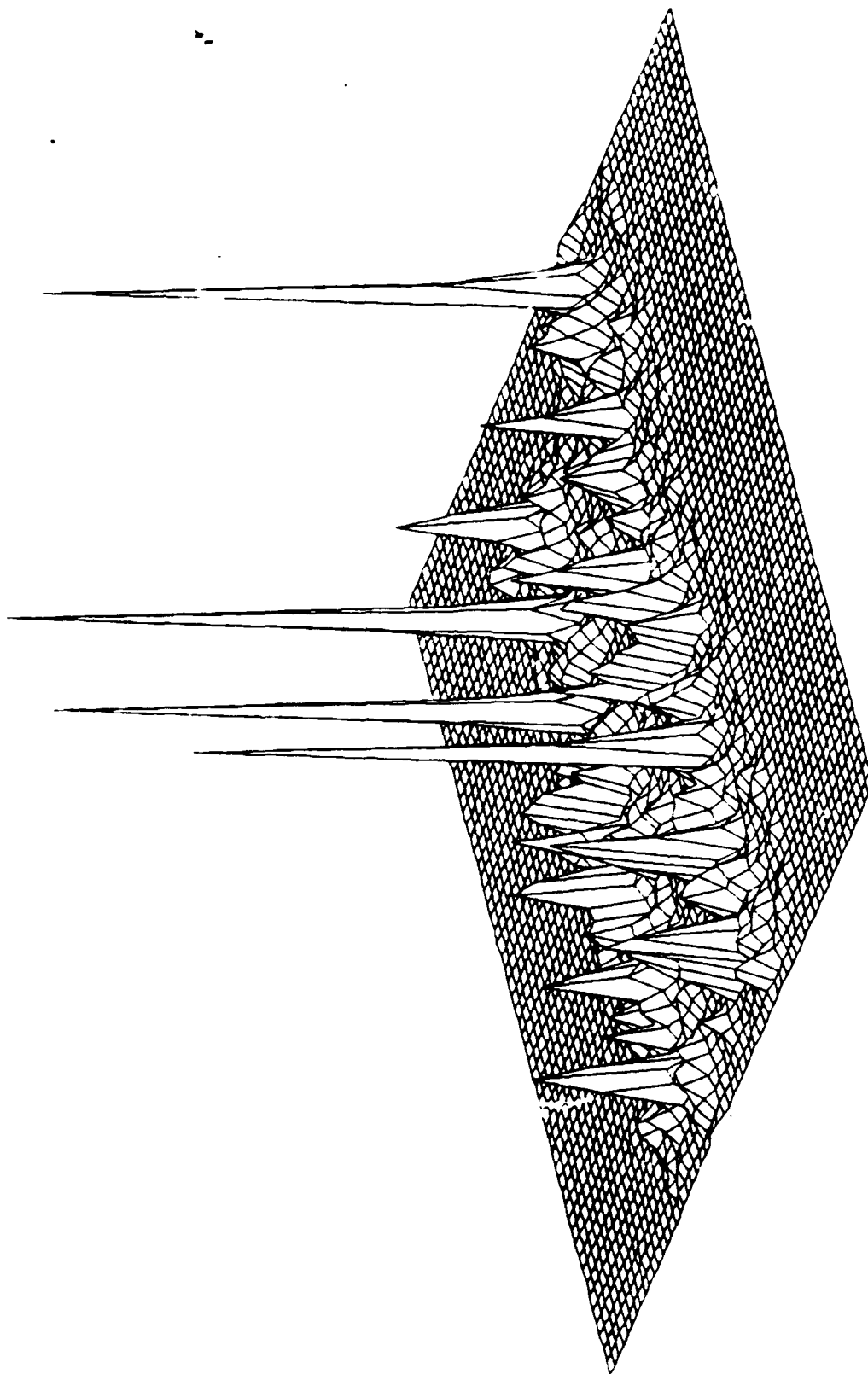


Figure 23. Correlations of the letter "O" with OKLAHOMA SOONERS. Remember that the text is inverted in the correlation plane.

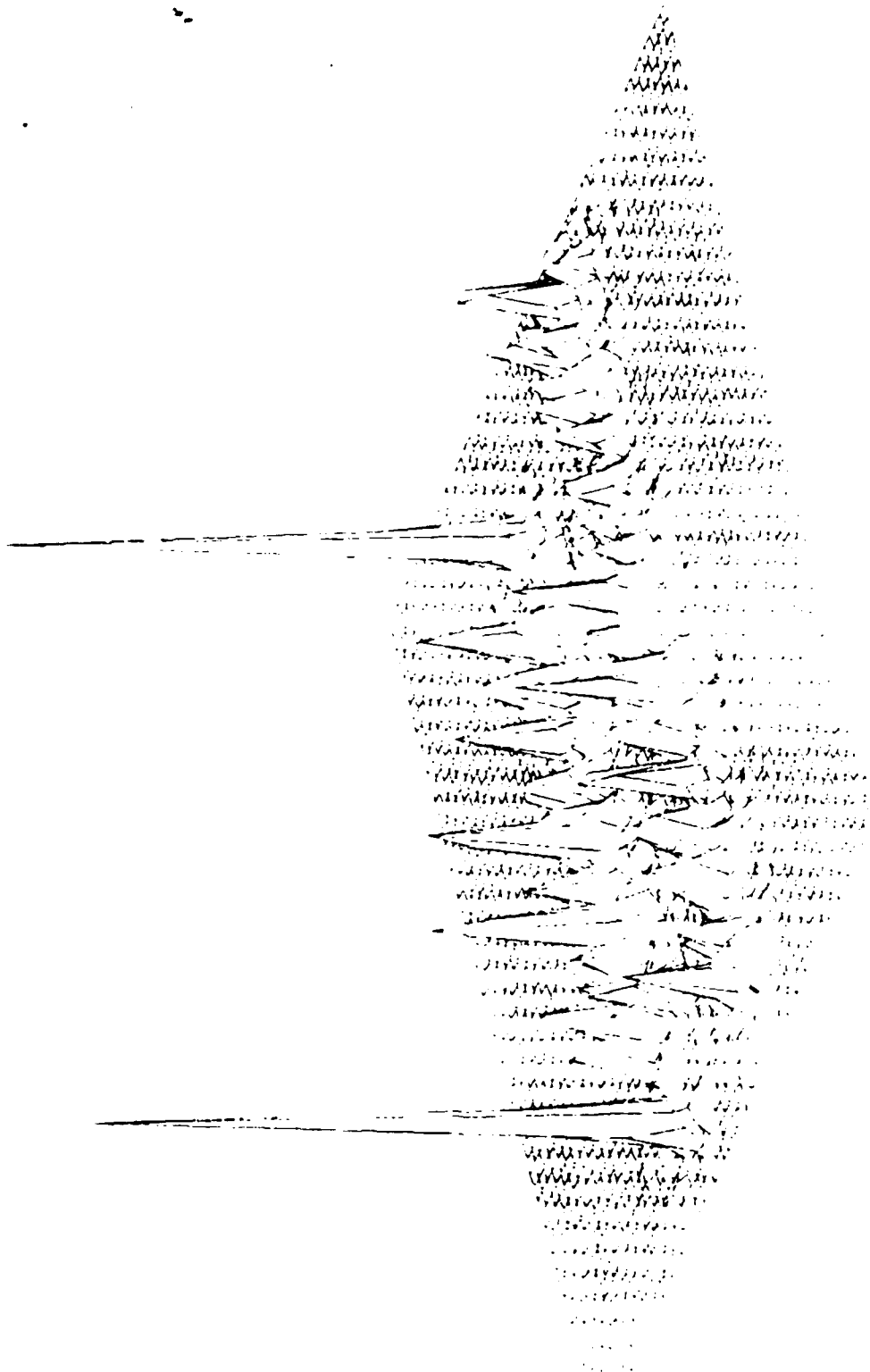


Figure 24. Correlations of the letter "S" with OKLAHOMA SOONERS.

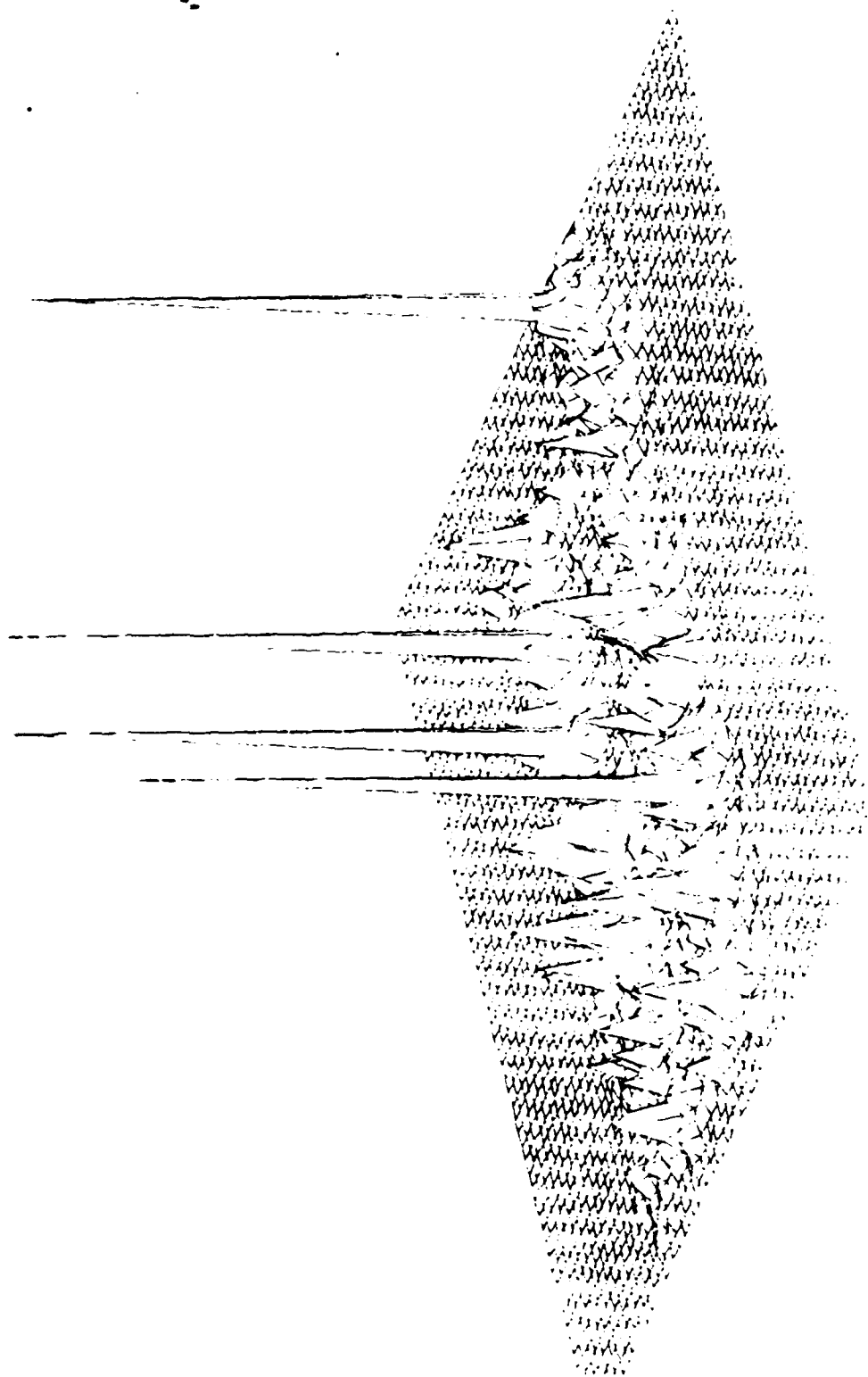


Figure 25. Effects of a high-pass filtering operation on the correlation of the letter "O" with OKLAHOMA SOONERS.

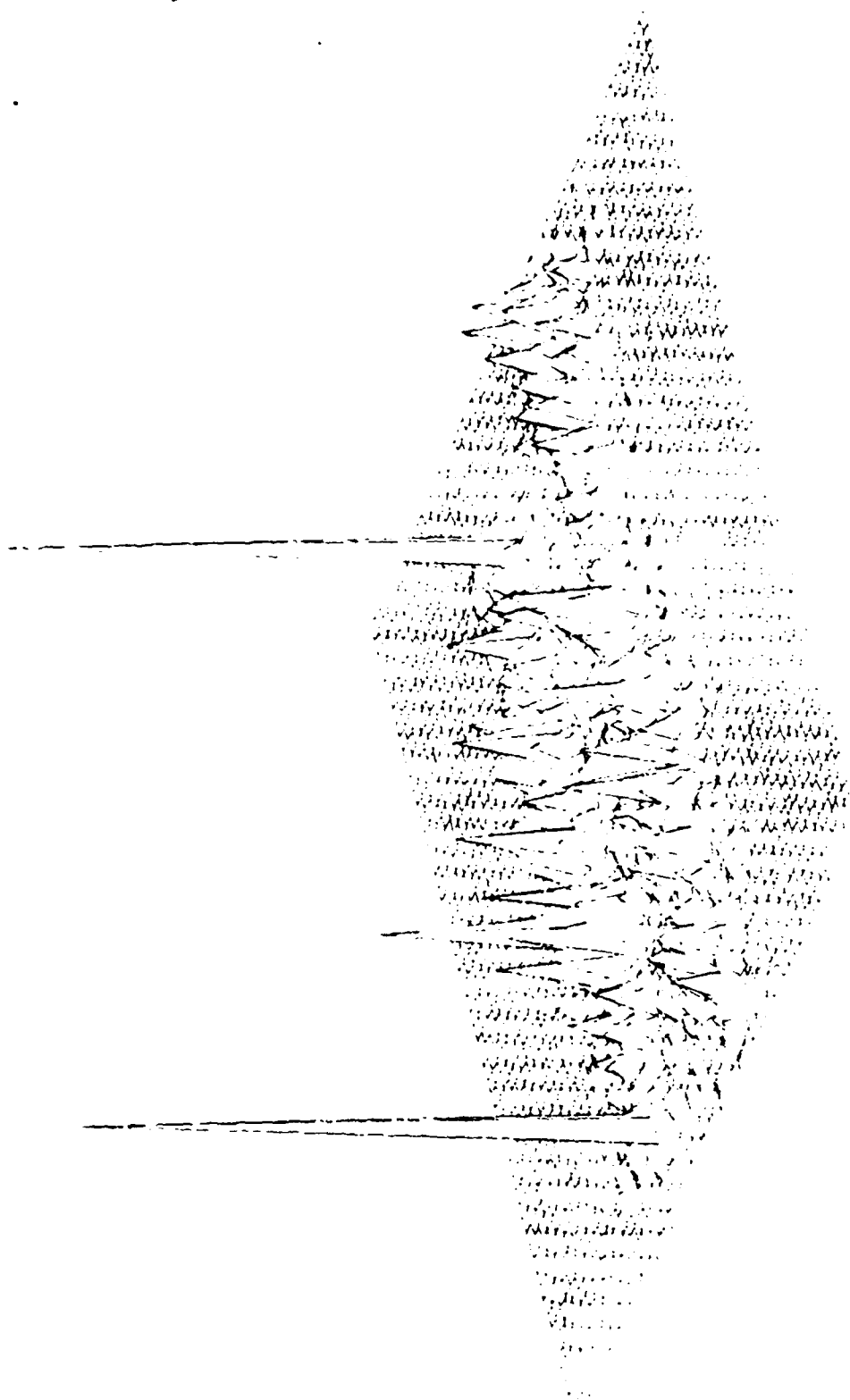


Figure 26. Effects of a high-pass filtering operation on the correlation of the letter "S" with OKLAHOMA SOONERS.

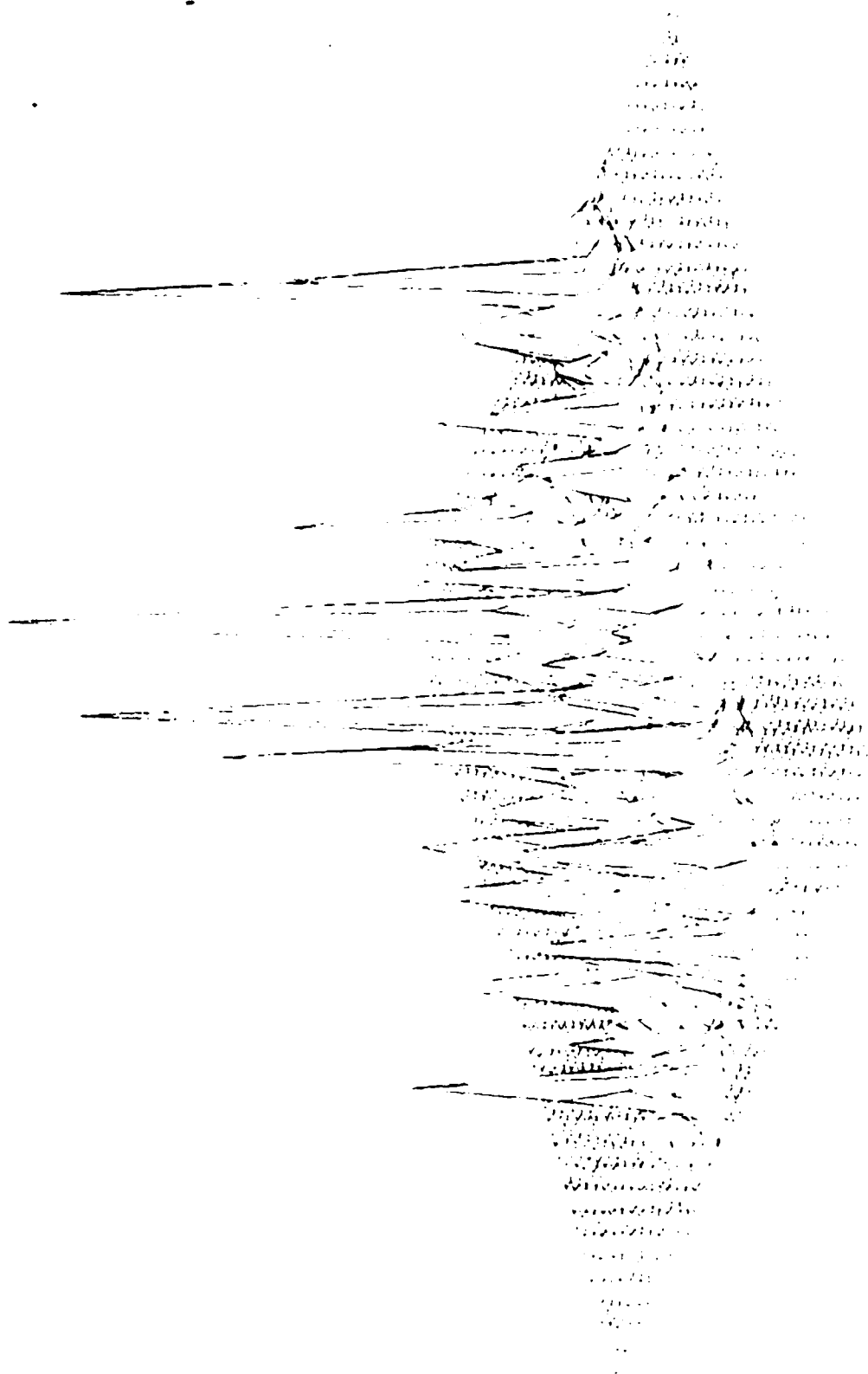


Figure 27. Effects of a low-pass filtering operation on the correlation of the letter "O" with OKLAHOMA SOONERS.

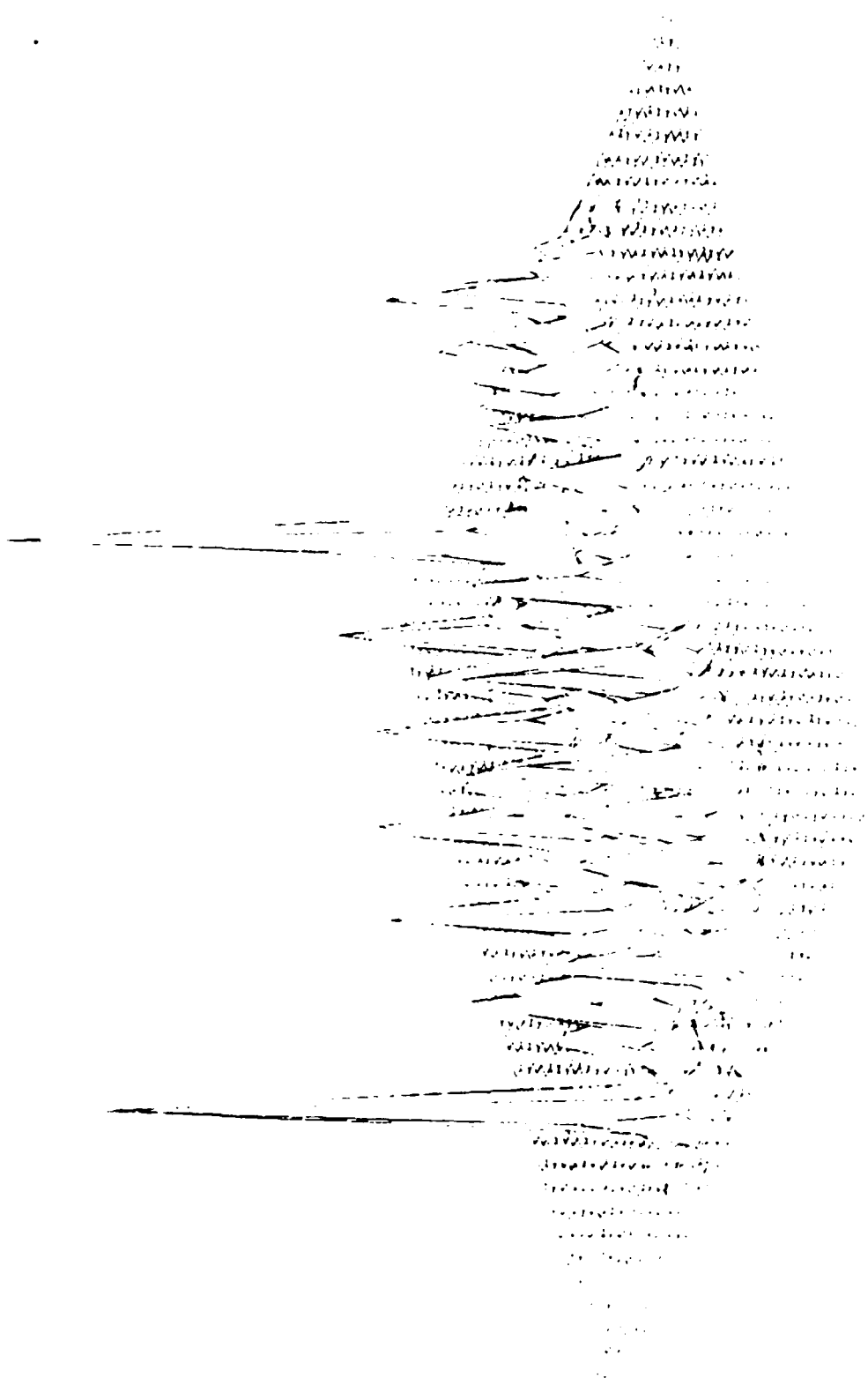


Figure 28. Effects of a low-pass filtering operation on the correlation of the letter "S" with OKLAHOMA SOONERS.

Birefringent Axes Vs. Contrast Voltage

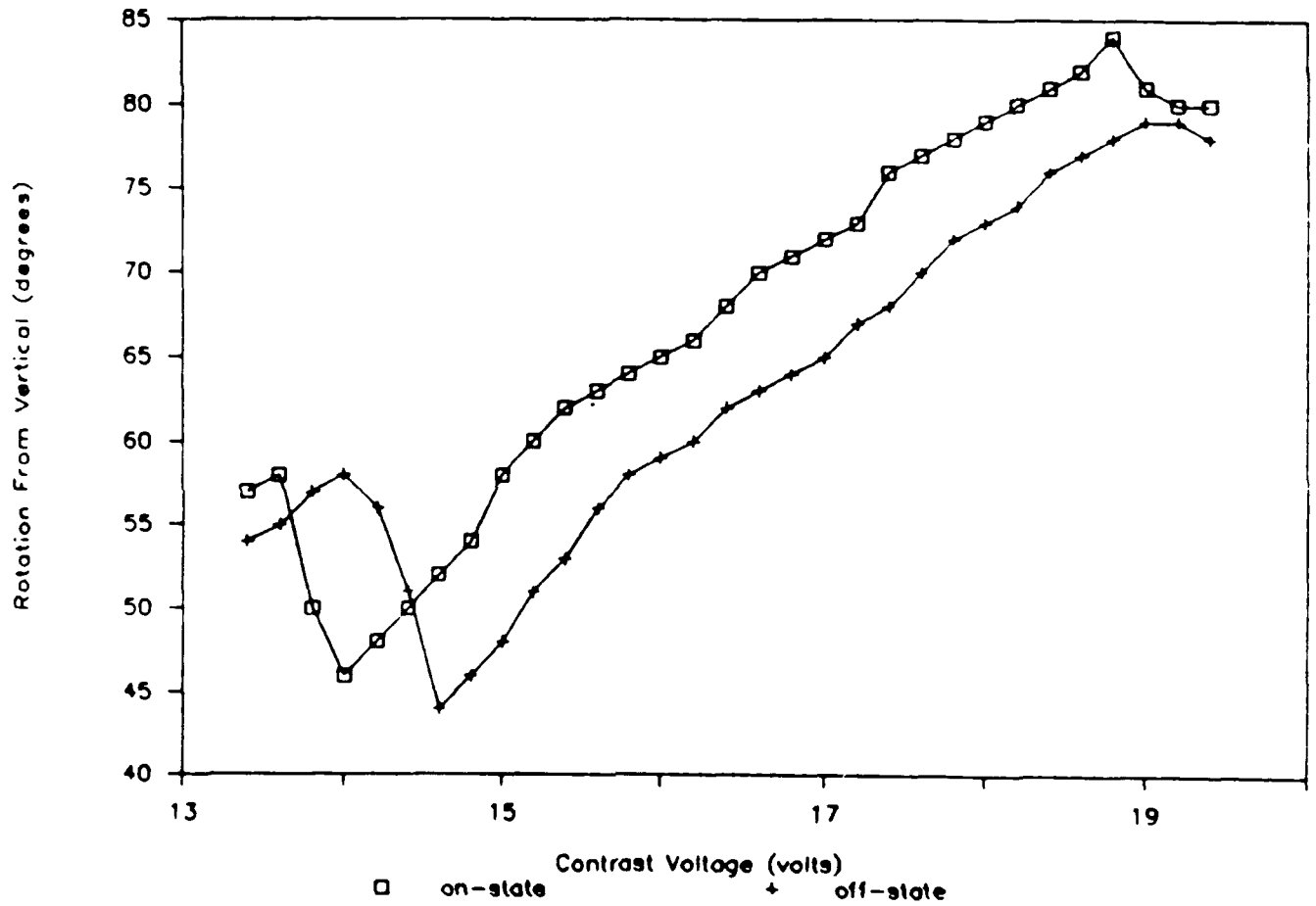


Figure 29. LCTV birefringent axis angles as a function of contrast voltage.

Polarization Angles Vs Contrast Voltage

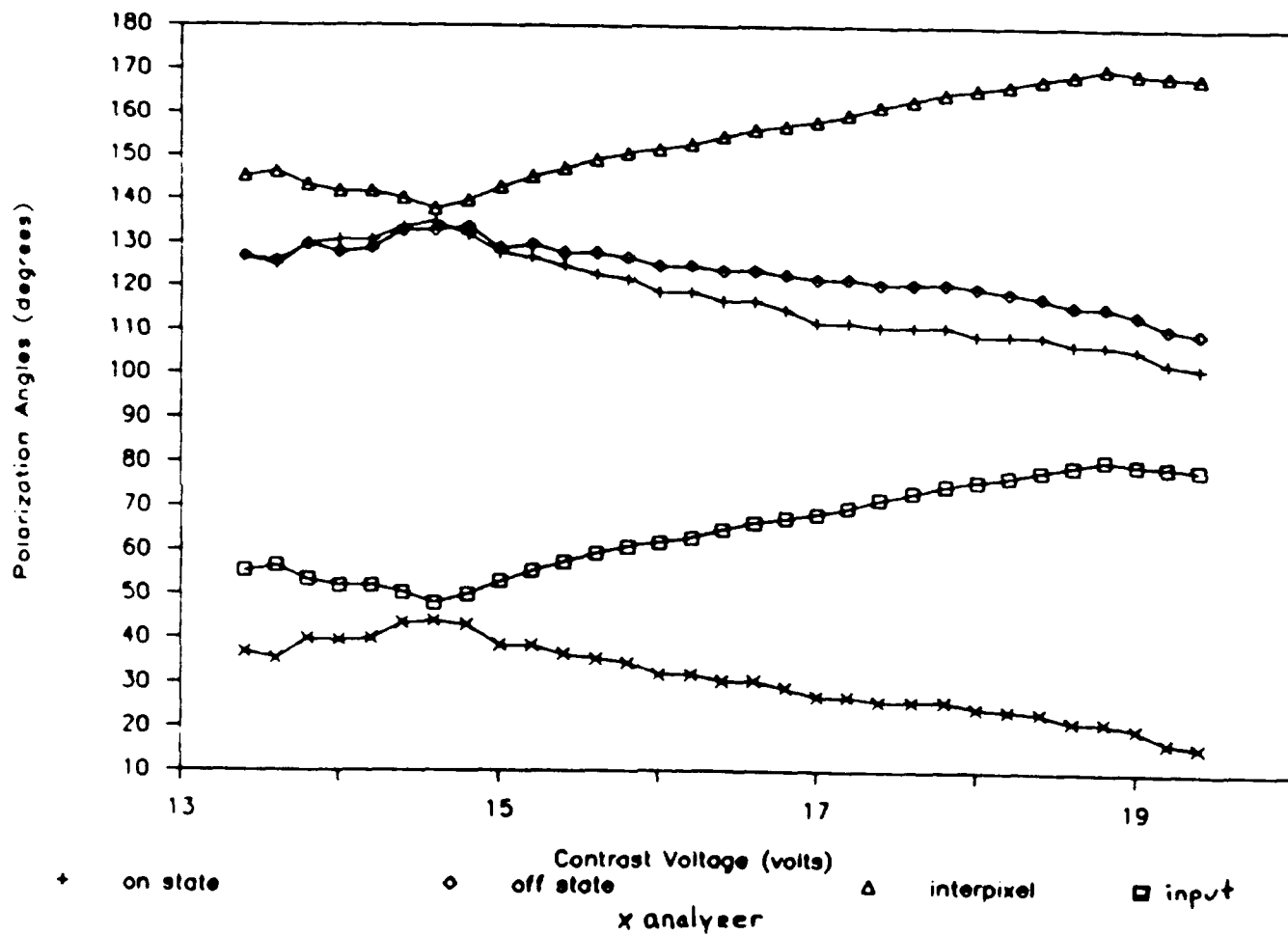
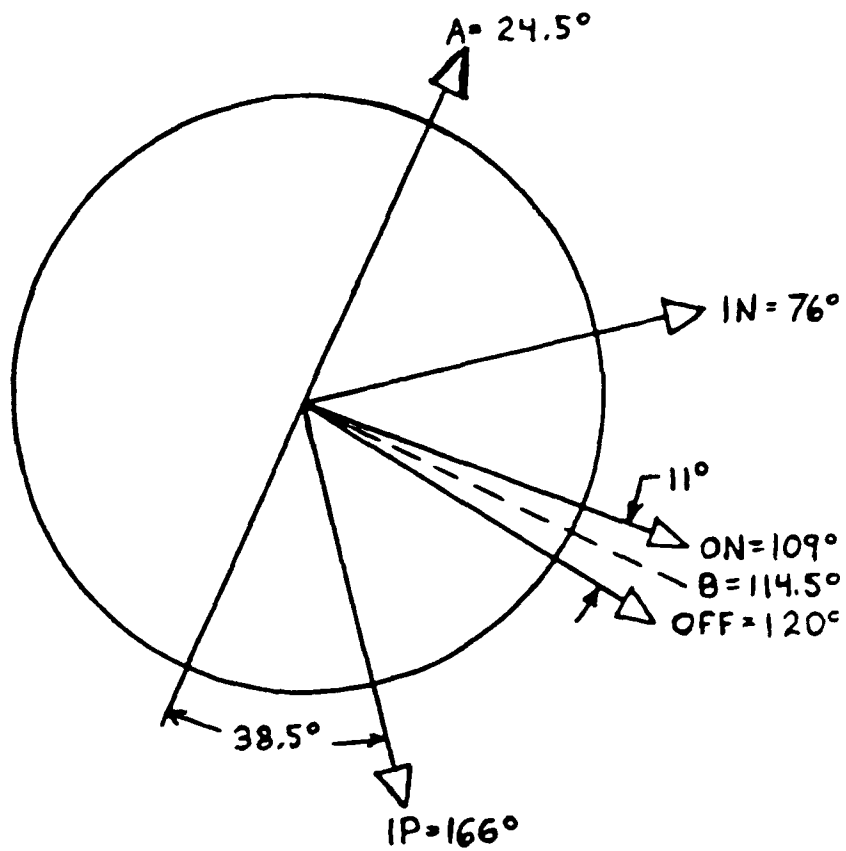


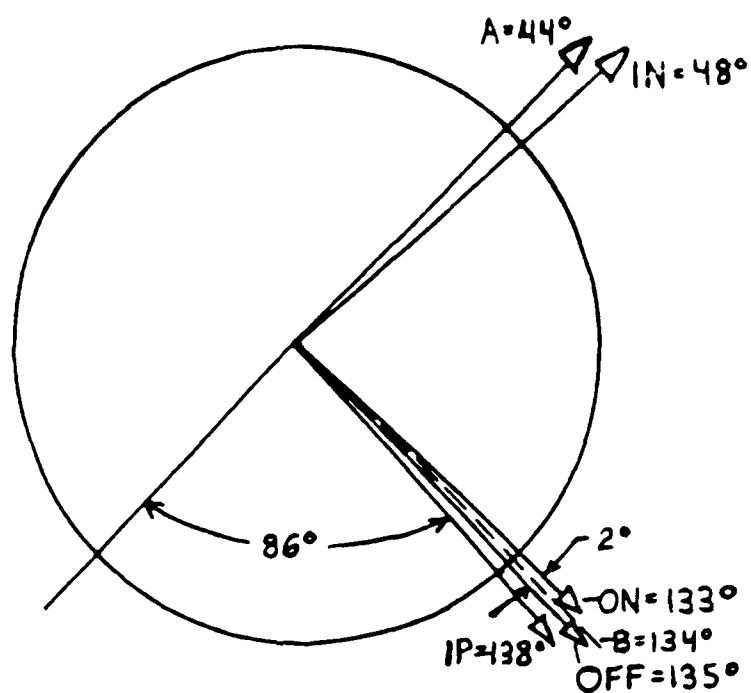
Figure 30. LCTV polarization angles as a function of contrast voltage.



IP - Interpixel
 A - Analyzer
 IN - Input

B - Bisector
 OFF - Off State Output
 ON - On State Output

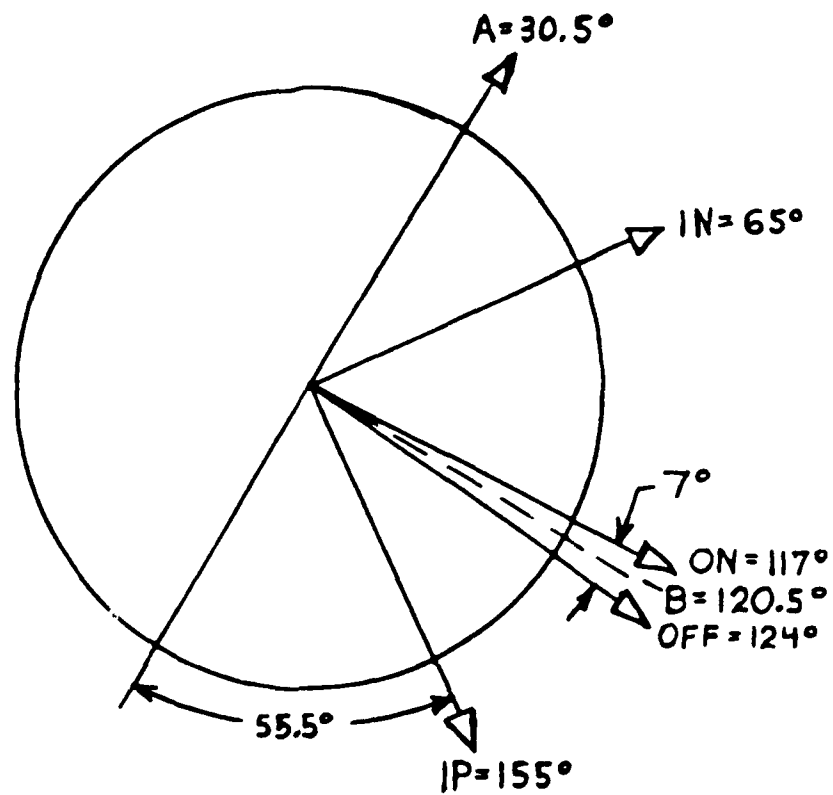
Figure 31. Vector diagram of LCTV polarization angles for maximum difference between on state and off state output polarization angles (contrast voltage used: 18 V).



IP - Interpixel
 A - Analyzer
 IN - Input

B - Bisector
 OFF - Off State Output
 ON - On State Output

Figure 32. Vector diagram of LCTV polarization angles for maximum blockage of interpixel light (contrast voltage used: 14.6 V).



IP - Interpixel
 A - Analyzer
 IN - Input

B - Bisector
 OFF - Off State Output
 ON - On State Output

Figure 33. Vector diagram of LCTV polarization angles for the compromise setting used for the experimental data runs (contrast voltage used: 16.4 V).

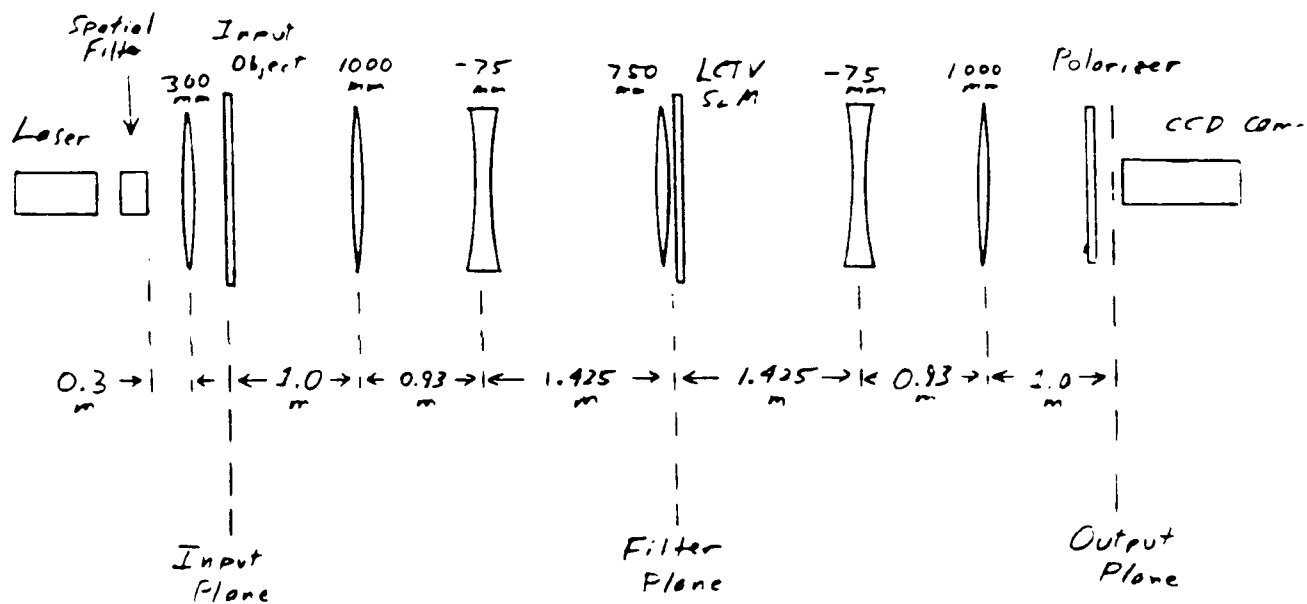


Figure 34. The test correlator configuration.

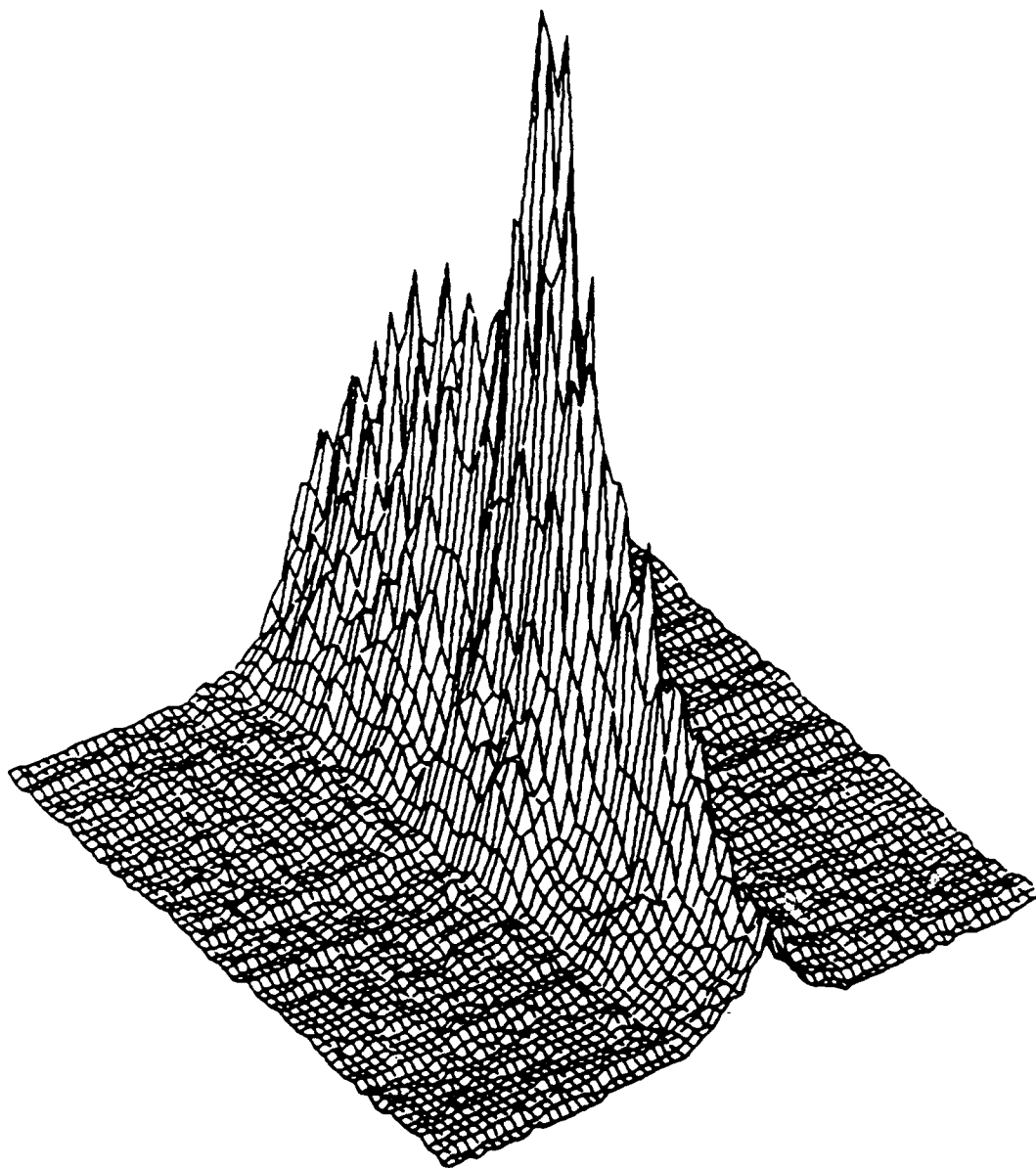


Figure 35. Experimental autocorrelation result using a LCTV BPOF:
Example No. 1. The relative correlation peak intensity is 220.

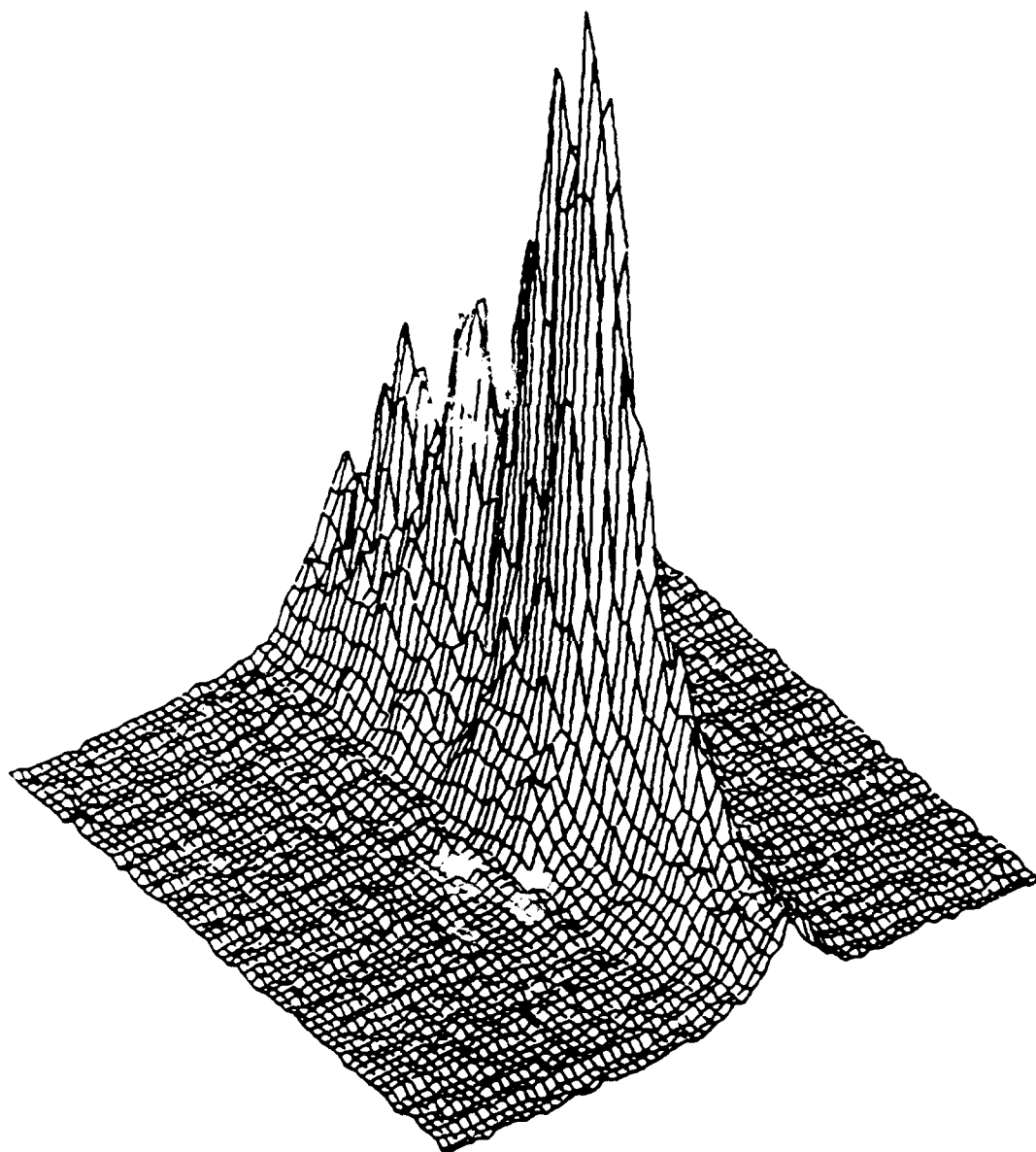


Figure 36. Experimental autocorrelation result using a LCTV BPOF: Example No. 2. The relative correlation peak intensity is 255.

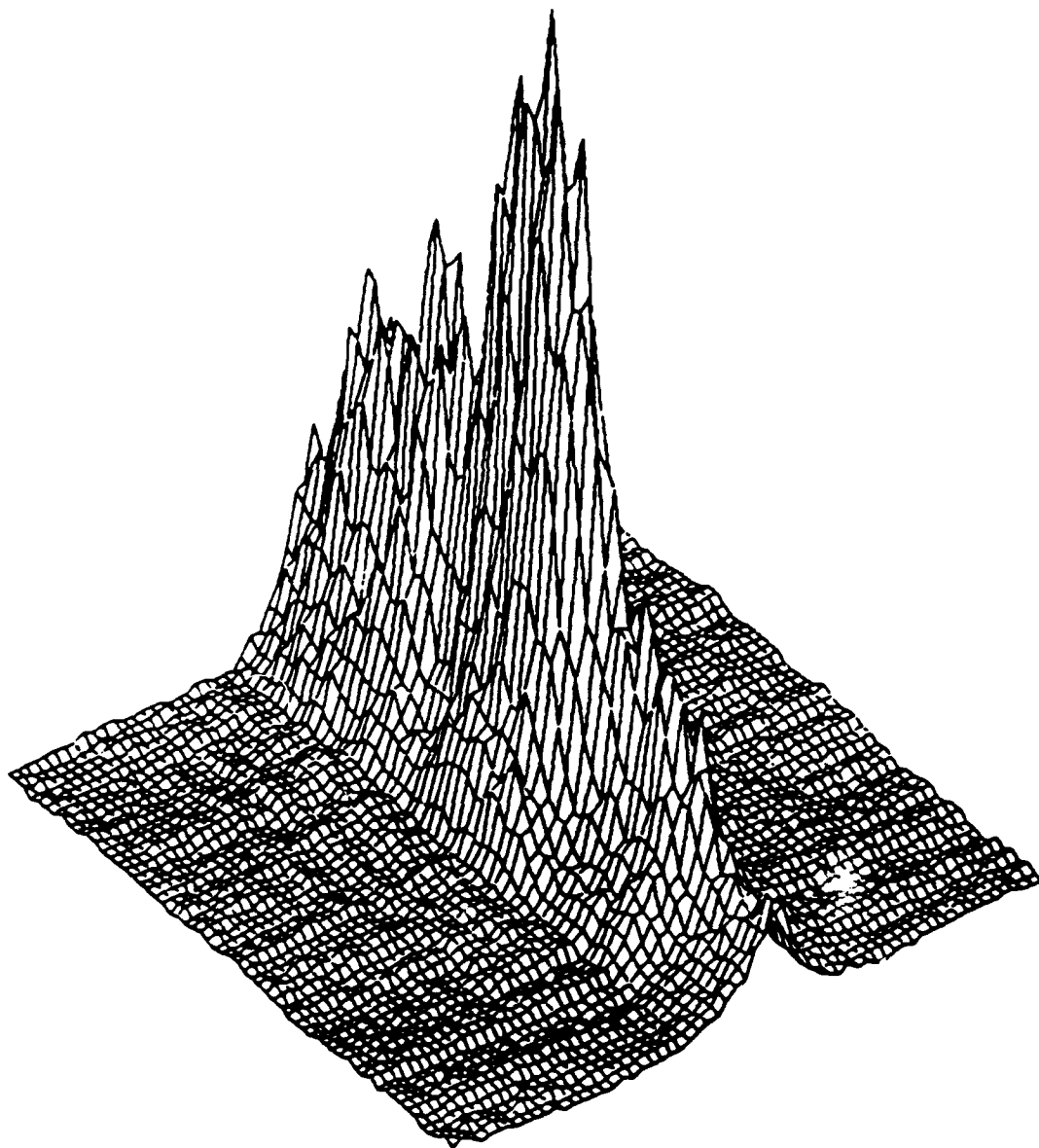


Figure 37. Experimental autocorrelation result using a LCTV BPOF: Example No. 3. The relative correlation peak intensity is 234.

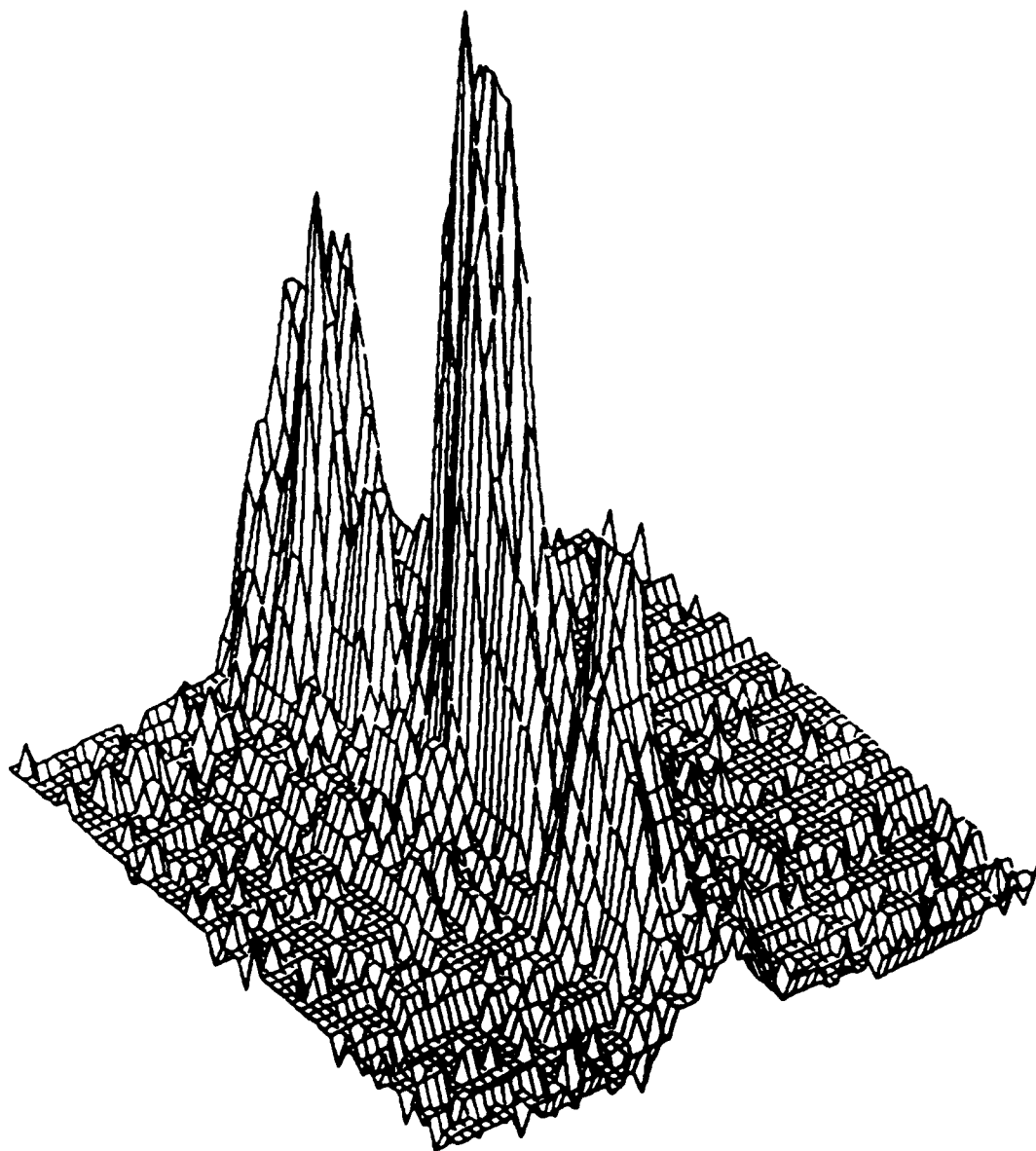


Figure 38. Same as Figures 35 - 37, but with added high-pass filtering. The relative correlation peak intensity is 86.

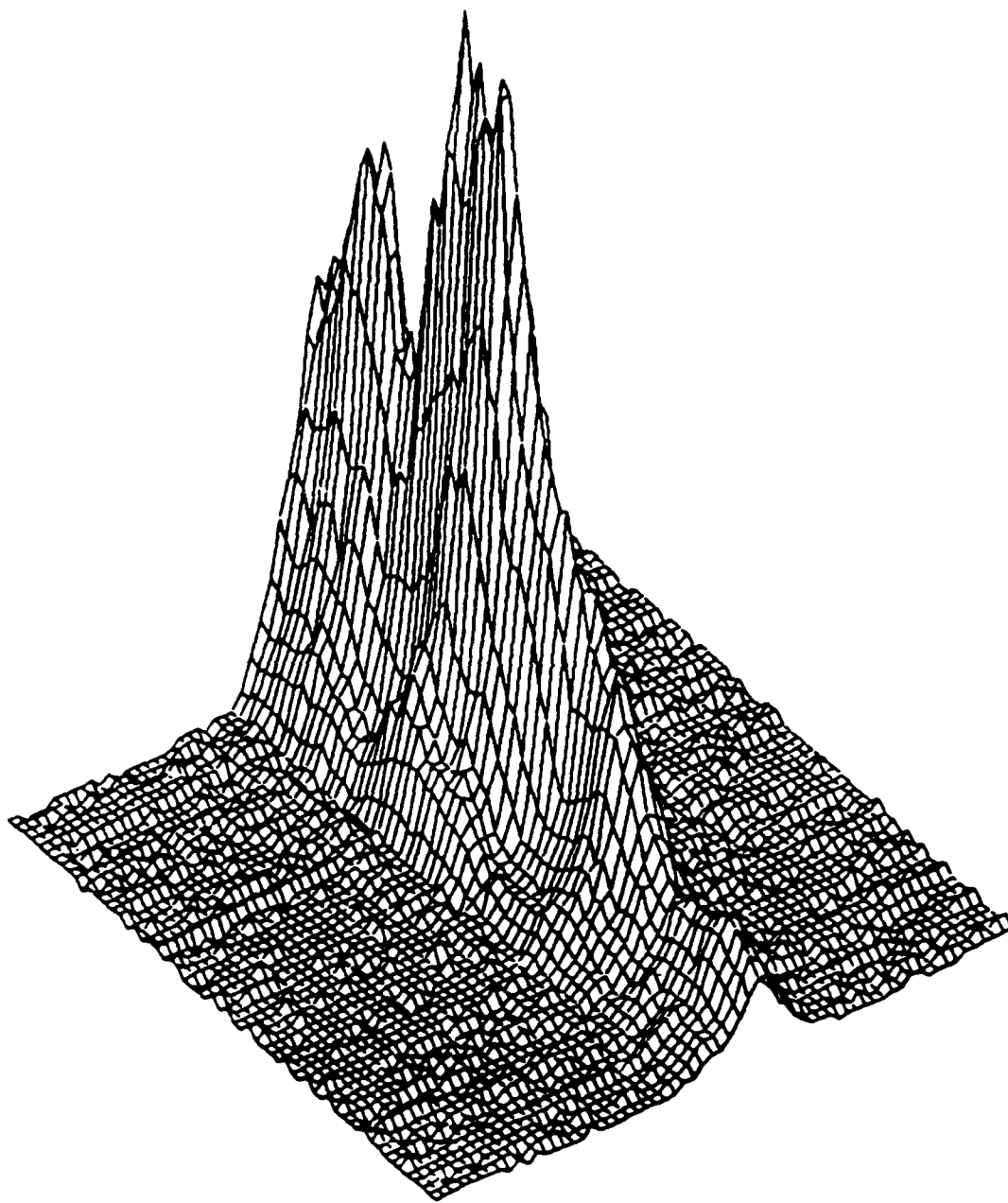


Figure 39. Same as Figures 35 - 37, but with added low-pass filtering. The relative correlation peak intensity is 165.

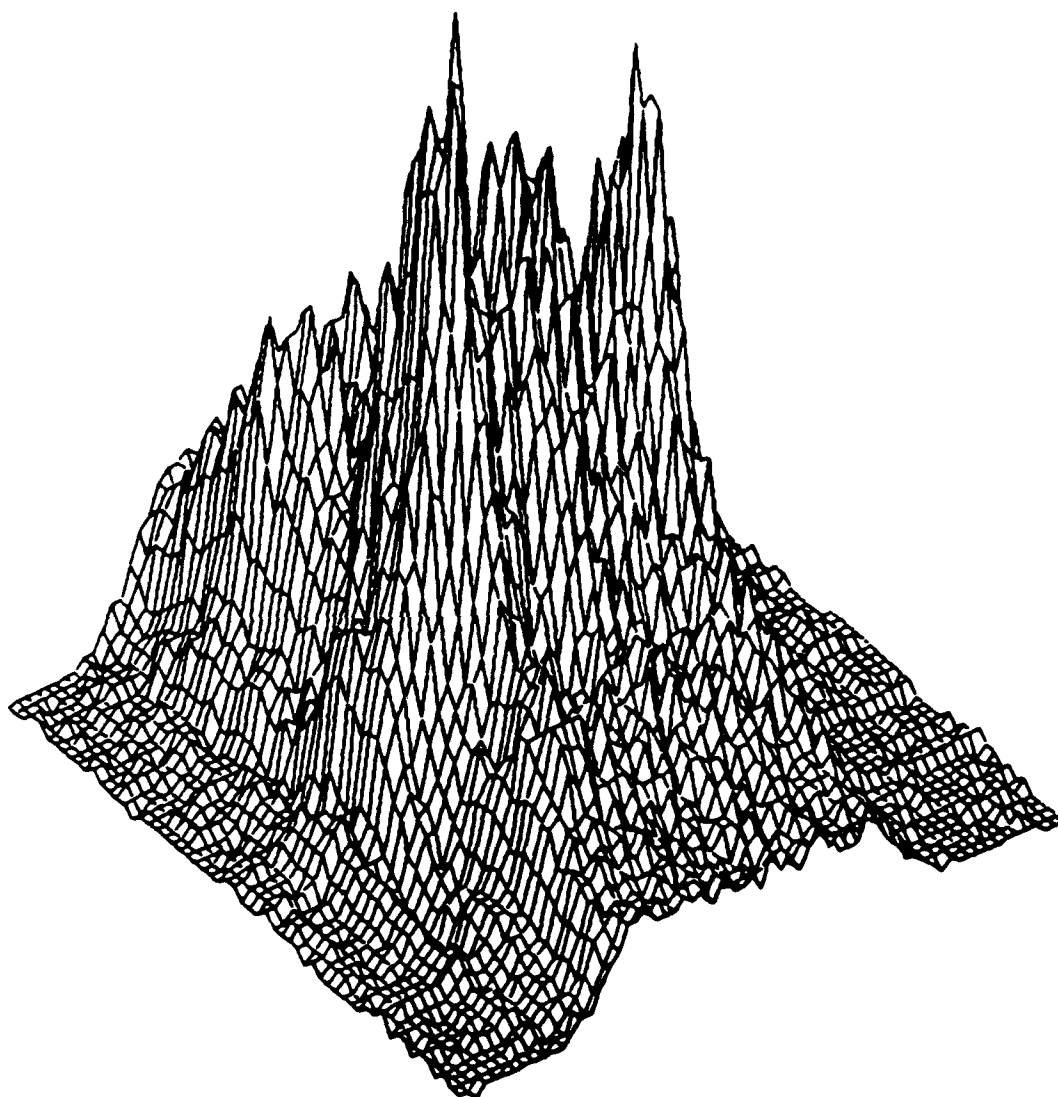


Figure 40. Example of a cross-correlation of a rectangle pattern with a different rectangle. The higher correlation peak has a relative intensity of 144.

**USAF-UES MINI GRANT
RESEARCH PROGRAM**

Sponsored By

**AIR FORCE OFFICE OF
SCIENTIFIC RESEARCH**

Conducted By

Universal Energy Systems, Inc.

FINAL REPORT

**Computer Modeling of GaAs/AlGaAs
MQW Devices for Optical Properties**

Prepared by:	Khaja F. Subhani, Ph. D.
University:	Formerly Associated with Lawrence Technological University Electrical Engineering Dept. -- Southfield, Michigan 48075
Current Address:	University of Alabama Physics and Astronomy Dept. 202 Gallalee Hall Tuscaloosa, AL 35478
Research Location:	Rome Laboratory/OCPB Griffiss AFB, NY 13441-5700
RL/OCPB Researcher:	Richard Michalak, Ph. D.
Date:	February 5, 1992
Contract Number:	F49620-88-C-0053/SB5881-0378
Purchase Order No:	S-210-10MG-107

ACKNOWLEDGEMENTS

Due to an unavoidable circumstance, I could not complete this task on time. However, a request for an extension of time was granted, graciously, by all the parties concerned for which I am deeply indebted specially to Air Force Systems and Command and the Air Force Office of Scientific Research in general, and Dr. Richard Michalak of Photonic Laboratory, GAB, NY, in particular, for his understanding and support. I have special words of thanks for Dr. Michael Parker of Photonic Laboratory, GAB, NY, for skillful proof reading and editorial comments. Thanks are also due to Universal Energy Systems, Inc for their concern and directional aspect of the program. Melissa Tomlan, Minigrant secretary of UES, deserves special mention for her patience, perseverance and understanding. Understanding, cooperation and support of Director Melvin Janney of Lawrence Technological University, Southfield, Michigan is also gratefully acknowledged. I could not possibly have completed this work without patience, perseverance and support of my family. Above all, I am thankful to God for this cheerful experience.

Chapter I
Computer Modeling of GaAs/AlGaAs
MQW Devices for Optical Properties

Introduction

The objective of this research is to develop a heterostructure device model leading to an optical gate for optical computing systems. One form of optical element is similar to an injected laser diode, and is thus reasonably compatible with the size of optical computer processing elements, integrated optic waveguide and optical fiber interconnection. Multiple Quantum Well (MQW) devices have appeared as logic gates to perform entire range of digital logic functions, similar to conventional electronics building blocks. [1]

It is well known that the introduction of quantum wells (with a thickness $L_z \sim 100\text{\AA}$) in the active layer of a double heterostructures GaAs/AlGaAs drastically improves its performance [6-7]. In these structures the carrier motion normal to the active layers is restricted. As a result, the kinetic energy of the carriers moving in the normal

direction are quantized into discrete energy levels similar to the well-known quantum mechanical problem of the one dimensional potential well, and hence the name quantum well (QW) lasers. When the thickness of the active region becomes comparable to the de Broglie wavelength ($\lambda \sim h/p$), quantum mechanical effects are expected to occur. These effects are observed in the absorption and emission characteristics and transport characteristics including phenomena such as tunneling. The optical characteristics of semiconducting QW double heterostructures were initially studied by Dingle et al [2]. Since then extensive work on GaAs/AlGaAs QW structures has been done by Holonyak et al[9,10], Tsang[11,12] and Hersee et al [13]. In recent years, GaAs/AlGaAs quantum well lasers have been of great interest because of the special properties associated with the two-dimensional-like nature. However, due to the great number of parameters involved, the solution to the problem of device optimization is not simple, and the discussion about the choice of the well characteristics in order to obtain the lowest possible threshold current is still open.

In the following chapters analytical models are presented to characterize the MQW ridge structure. These include, quantum confined

carrier energy levels, carrier gain, leakage current, nonradiative recombination current, gain confinement factor and finally the threshold current models. In chapter III calculation technique and optimal device performance model for minimum threshold current is described. Following this, optical gain calculation is presented in this chapter, including the technique to optimize the structural parameters for device performance models. Conclusions are presented in chapter IV. Finally computer models are listed in chapter V.

Chapter II

Theoretical Models for GaAs/AlGaAs

MQWs-Application to Ridge Structure

II.1 Single Quantum Well Energy Levels

The quantum size effect which takes place when one of the sample dimensions is on the order of the de Broglie wavelength of the carriers is now well known [10,14]. The electrons and holes experience quantization in the confined direction and their energy can be written as[15]:

$$E = \frac{\hbar^2(k_x^2 + k_y^2)}{2 \cdot m_{e,h}^*} + E_i, \quad (1)$$

where E_i is the i th confined particle energy, $m_{e,h}^*$ is the electron or hole effective mass, \hbar is Plank's constant, and k_x and k_y are the x and y components of the crystal momentum whose quasi-continuous values form parabolic subbands minimum energy is the quantized energy level E_i . The origin of the energy is taken at the maximum of the bulk valence band of GaAs. Assuming a symmetric square well with finite barrier height, the energy of the quantized levels can be found for MQW structure by solving the eigenvalue equations:

$$A \cdot \tan(A \cdot \frac{L_z}{2}) = B, \quad \text{for even solutions;} \quad (2a)$$

$$A \cdot \cot\left(A \cdot \frac{L_z}{2}\right) = -B, \quad \text{for odd solutions.} \quad (2b)$$

where

$$A = \left\{ \frac{2 \cdot m_a^* \cdot E_i}{\hbar^2} \right\}^{1/2} \quad (3a)$$

and

$$B = \left\{ \frac{2 \cdot m_b^* \cdot (V - E_i)}{\hbar^2} \right\}^{1/2}. \quad (3b)$$

In the above equations m_a^* is the effective mass of carriers in the well, m_b^* is the effective mass of the carriers in the barrier, V is the finite potential barrier, and L_z is width of the well.

Figure (1) shows schematically the energy levels E_i of the electrons confined in a quantum well. The confine particle energy levels (E_i) are denoted by E_{1c}, E_{2c}, E_{3c} for electrons, $E_{1hh}, E_{2hh}, E_{3hh}$ for heavy holes and $E_{1lh}, E_{2lh}, E_{3lh}$ for light holes. Thus in a QW structure the energy of the emitted photons can be varied simply by varying the well width L_z . In the GaAs/GaAlAs system, the mass is determined as a function of the aluminum composition x of the layer as [16]:

$$m_e = (0.067 + 0.0835x)m_o, \quad (4a)$$

$$m_h = (0.48 + 0.31x)m_o, \quad (4b)$$

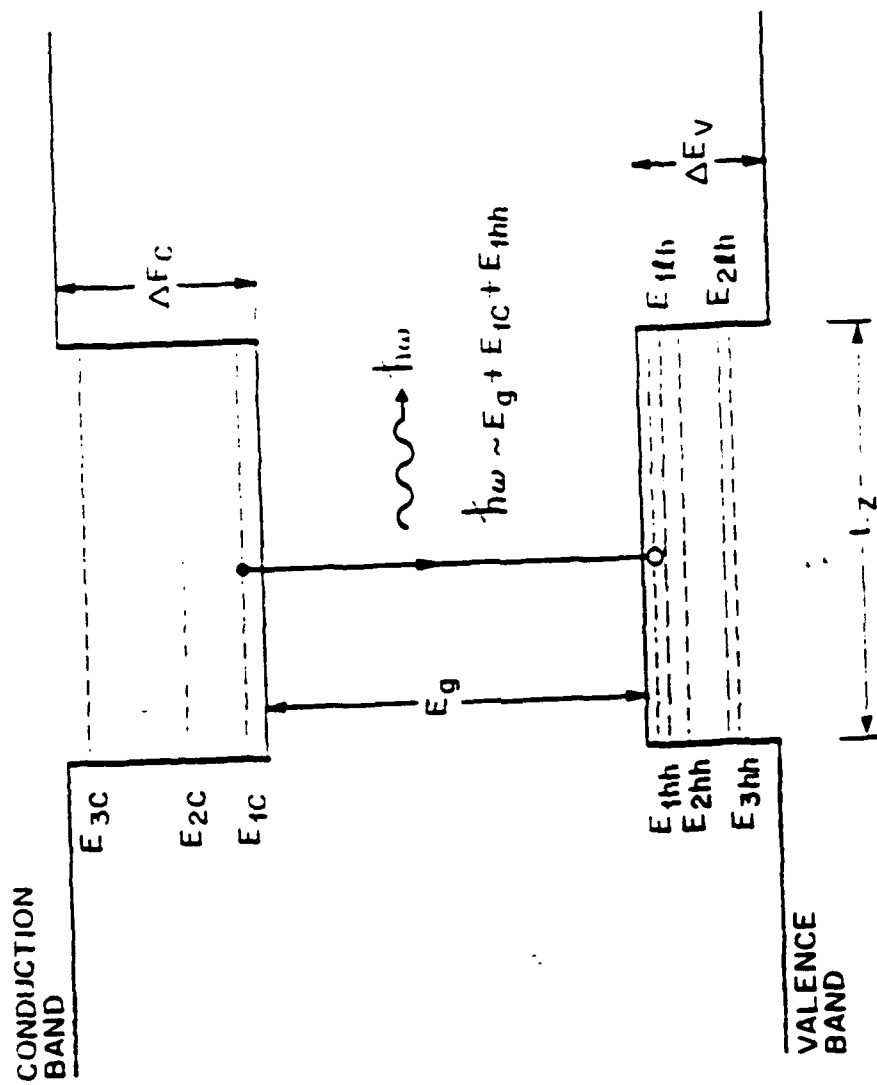


Fig.(1): Energy levels in a QW structure.

where m_o is the free electron mass (the light holes are neglected). The total hetero barrier height is defined as:

$$\Delta E_{go} = E_g(\text{barrier}) - E_g(\text{well})$$

where

$$E_{go} = 1.424 + 1.247x, \quad \text{if } x \leq 0.45 \quad (5a)$$

$$E_{go} = 1.424 + 1.247x + 1.147(x - 0.45)^2, \quad \text{if } x \geq 0.45 \quad (5b)$$

Although the values of the energy discontinuities of the conduction and valence bands are largely discussed [17], this model uses barrier height $V_e = 0.58 \cdot \Delta E_{go}$ for electron, $V_h = 0.32 \cdot \Delta E_{go}$ for holes, and $V = V_e + V_h$ [8]. Equation (2a) and (2b) then allow the determination of quantized levels for electrons and holes. The density of states, constant within one subband, is given by [9]:

$$g_{e,h} = \frac{m_{e,h}^*}{\pi \hbar^2 L_z} \quad (6)$$

This model uses the classical Fermi statistics, assuming that electrons and holes are in equilibrium in the well and that all carriers in the same band are characterized by the same quasi-Fermi level. They

are determined using the condition of electrical neutrality:

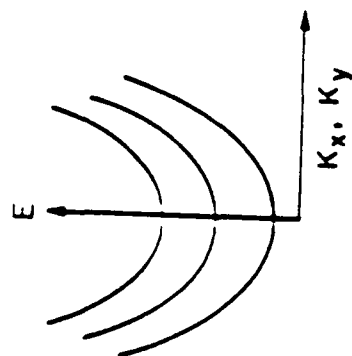
$$n = n_a + p \quad (7)$$

where n_a is the p-type impurity concentration, if any in the QW layers. A comparison of equation (6) with a regular three-dimensional density of states model shows that the density of states in the QW is independent of carrier energy and temperature. The modification of density of states in a QW is sketched in figure (2).

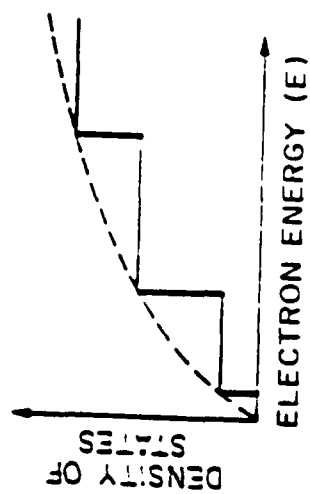
The electron density n is obtained by integrating over k_x and k_y in the x and y directions and by summing the contribution of each discrete subband. Hence, after calculations, n can be expressed as:

$$n = \frac{m_e^* \cdot kT}{\pi \cdot \hbar^2 \cdot L_z} \sum_{i=1}^M \left\{ \ln \left(1 + \exp \left(\frac{E_{fc} - E_i}{k \cdot T} \right) \right) \right\} \quad (8)$$

where L_z is the well width, M is the the number of quantized levels in the well, E_i is the minimum of the i th subband, and E_{fc} is the Fermi level for electrons. A similar expression can be written for the valence band hole density. We consider only the heavy holes since the density of states of the light hole band is much smaller than that of the heavy hole band [18,19]. For given values of n and n_a , (7), (8), and the similar equation giving p allow determination of the Fermi levels E_{fc} and E_{fv} for both the conduction and the valence bands.



(a)



(b)

Fig.(2): (a) Energy versus wavevector for each subband;
(b) Schematic representation of the density of states in a QW.

II.2 Gain in GaAs/AlGaAs MQWs

To examine the operating characteristics of MQW structures theoretically, one needs to study the gain carrier distribution as a function of photon energy and its dependence on the injection current. The gain distribution under on (lasing) condition was reported to be slightly inhomogeneous (in the QWs) for an undoped active region and to be fairly homogeneous when the active region was heavily doped [20,22]. The so called "band tail" was also well explained by the intra-band relaxation process, although the parabolic (no tail) states density and wavenumber conservation rule at optical transition were assumed [23]. The intra-band relaxation process is thus one of the most important phenomena characterising the gain of the laser material. Furthermore, gain calculations needs a numerical integration when the intra-band relaxation time is taken into account, as was shown in [20], [22], or [23].

A. Analytical Model for MQW Devices

The device structure shown in figure (3) is used for modeling the gain of MQW structures. The optical gain in QW structures with dif-

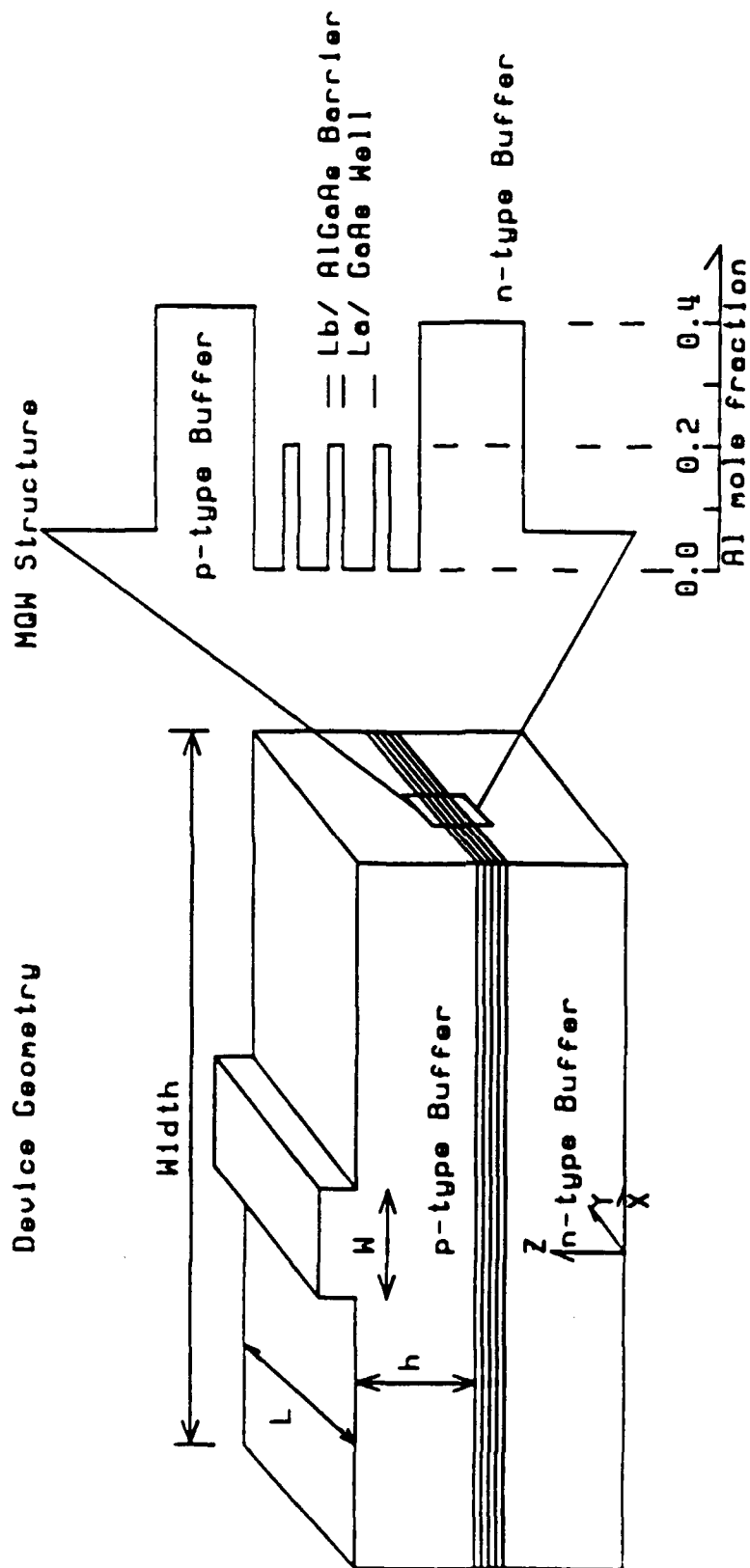


Figure (3): Device Geometry

ferent well widths has been calculated using an intraband relaxation model [18,23]. Detail expression for this model is shown in the following. However, this model has been chosen on one hand because the bell-like spectral shape of the gain curves [18] rules out the k-selection rule model, and on the other hand because the reported experimental differential gain values [19-24] are higher than those calculated by the no-selection rule model. In addition, the gain calculated using the chosen model agrees well with experiment reported in the literature[6]. According to the intraband relaxation model [18,23], the optical gain, $g(E)$, is calculated by the following expression:

$$g(E) = \text{const} \cdot \sum_n^{m-1} \int_{E_{cn}+E_{vn}+E_{go}}^{\infty} \left\{ \rho_{red} |M|^2 \cdot \frac{(f_c - f_v)\hbar/\tau_{in}}{(E_{cv} - E)^2 + (\hbar/\tau_{in})^2} \cdot \frac{1}{E_{cv}} \right\} dE_{cv} \quad (9)$$

with

$$f_c = \left\{ 1 + \exp\left(\frac{m_r}{m_c} \cdot \frac{E_{cv} - E_{gn}}{kT} - \frac{E_{fc} - E_c}{kT}\right) \right\}^{-1} \quad (10)$$

$$f_v = \left\{ 1 + \exp\left(-\frac{m_r}{m_v} \cdot \frac{E_{cv} - E_{gn}}{kT} + \frac{E_{fv} - E_v}{kT}\right) \right\}^{-1} \quad (11)$$

$$|M|_{TE}^2 = \frac{3}{4} |M_o|^2 \left(1 + \frac{E_{cn}}{\epsilon_{cn}} \right) \quad (12)$$

$$|M|_{TM}^2 = \frac{3}{4}|M_o|^2 \left(1 - \frac{E_{cn}}{\epsilon_{cn}}\right) \quad (13)$$

$$\epsilon_{cn} = E_{cn} + \frac{m_r}{m_c} \left(E_{cv} - E_{cn} - E_{vn} - E_{go}\right) \quad (14)$$

$$|M_o|^2 = 1.33m_o E_{go} \quad (15)$$

$$\rho_{red} = \frac{1}{2\pi^2} \left\{ \frac{2m_r}{\hbar^2} \right\}^{3/2} \Delta E_1^{1/2} \text{Int} \left\{ \frac{E - E_{gn}}{\Delta E_1} \right\}^{1/2} \quad (16)$$

Finally the term

$$const = \frac{3\hbar e^2}{m_o^2 c \epsilon_o \bar{\mu}} \quad (17)$$

where E_{cn}, E_{vn} represent, respectively, the quantized carrier energies in the conduction and valence bands, $\Delta E_1 = E_{c1} + E_{v1}$ and $E_{gn} = E_{cn} + E_{vn} + E_{go}$ with E_{go} representing the band gap, m_r is the reduced mass ($1/m_r = 1/m_c + 1/m_v$), E_{cv} represents the transition energy, E corresponding to energy of emitted photon, and τ_{in} is the intraband relaxation time of the electron and hole pair. Other terms such as f_c and f_v represents Fermi-Dirac distribution functions for electrons and holes respectively, $|M|_{TE}^2$ and $|M|_{TM}^2$ represents transverse electric and transverse magnetic gain, $|M|_o^2$ is the dipole moment, ρ_{red} is the radiative recombination rate, and finally $\bar{\mu}$ is average refractive index along the z -direction of the QWs.

For a given carrier density (n), the conduction band quasi-fermi level (E_{fc}) is calculated by:

$$n = \frac{m_c^* kT}{\pi \hbar^2 L_z} \cdot \sum_{i=1}^M \ln \left\{ 1 + \exp \left(\frac{E_{fc} - E_{ci}}{kT} \right) \right\} \quad (18)$$

Neglecting the residual doping in the layers and considering the charge neutrality criterion:

$$n = p. \quad (19)$$

Furthermore, valence band quasi-fermi level (E_{fv}) is obtained by

$$p = \frac{m_v^* kT}{\pi \hbar^2 L_z} \cdot \sum_{i=1}^M \ln \left\{ 1 + \exp \left(\frac{E_{fv} - E_{ci}}{kT} \right) \right\} \quad (20)$$

Optical gain, $g(E)$, is plotted as a function of photon energy for a given injection current in figure (5). The maximum gain, (g_{max}), value versus energy (E) at a given injection (J) is plotted versus the injection current (J) in figure (6). It is expected that as the injection current increases, the staircase density of states in QW structure induces first a rapid gain increase and then a saturation profile at high injection, which is interrupted by the contribution from the second quantized level. The influence of this specific gain-current behavior on the threshold current will be examined later.

Quantum well Laser Spectrum

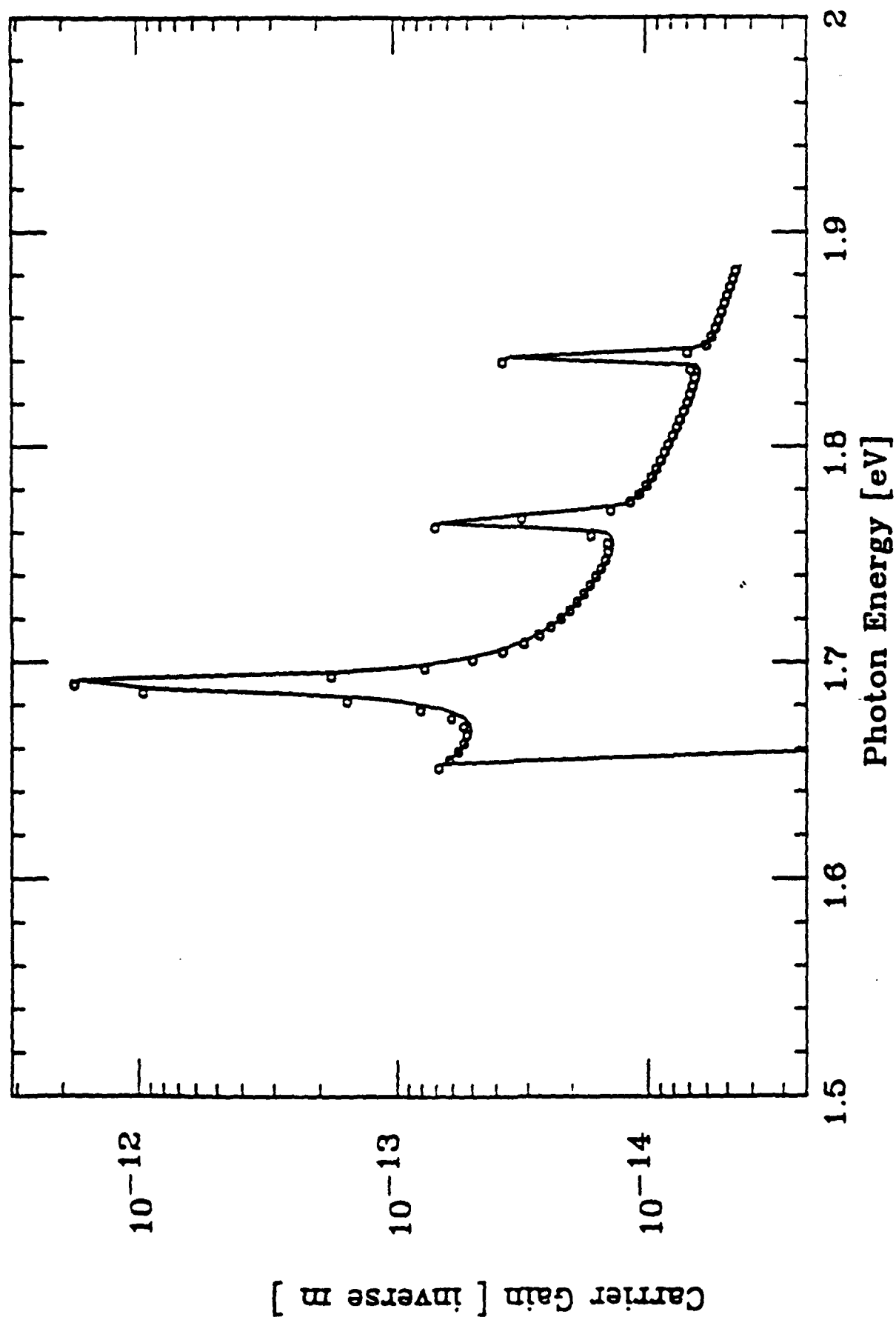


Figure (4)

CARRIER GAIN [m-1]: I_z=250.00[mA]

10⁻¹¹

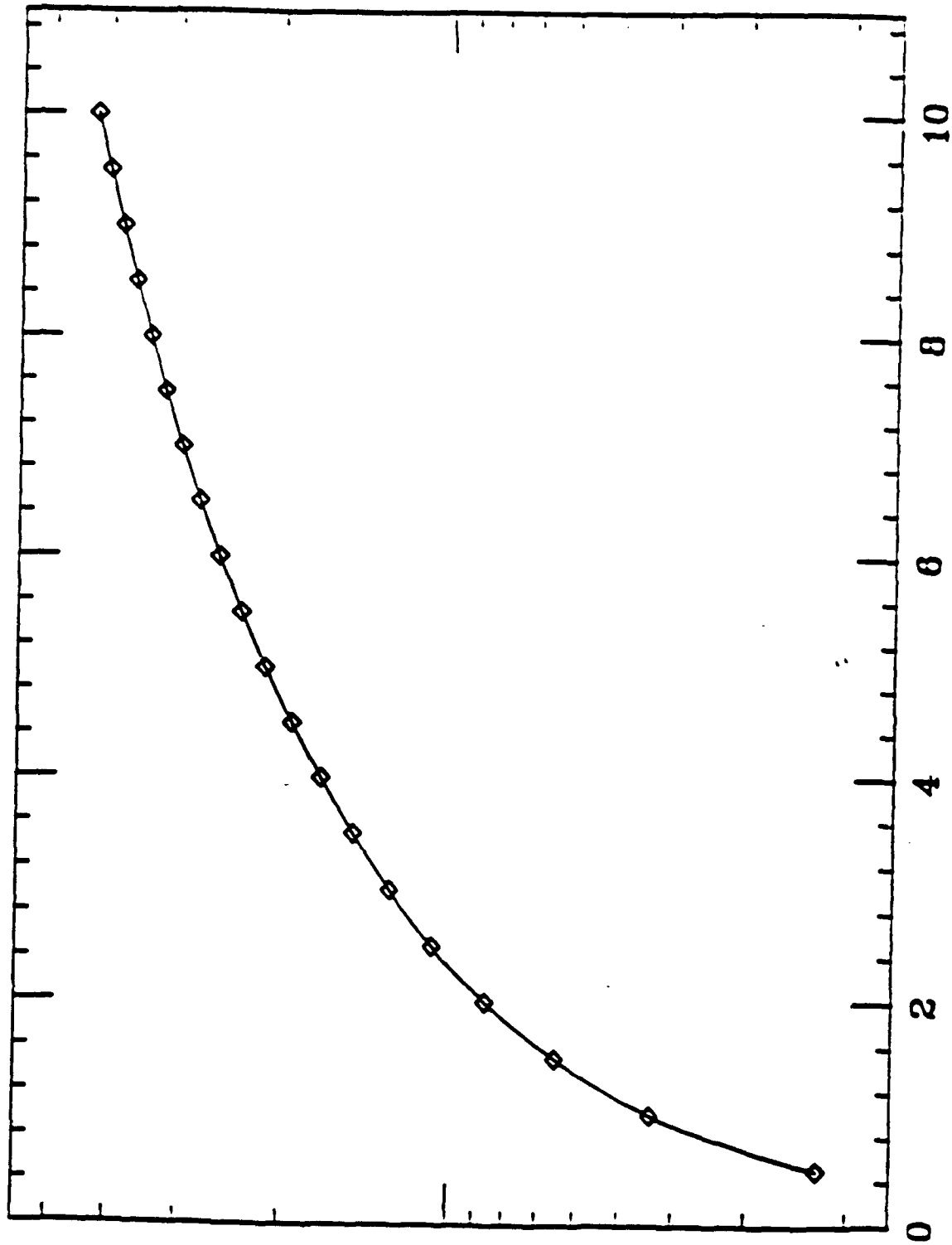


Figure (5)

4" IN DIAMETER FOR 0 WELLS

II.3 Leakage Current

A. Current in QW Ridge Structure

The calculation of lateral leakage current ($I_s + I_{diff}$) is performed using the same formulism as in the model developed by Joyce [25]. The two components I_s and I_{diff} of the lateral leakage current are illustrated in figure (6). The first component is an ohmic sheet current I_s (A/cm) (spreading current), which flows in a thickness h of the upper p -layer with sheet resistance $\Omega = \rho/h$; another component is a diffusion current I_{diff} (A/cm²), which flows in the active layer of thickness $d = M \cdot L_s$. In our calculation, Joyce's model is adopted to the QW structure by introducing the gain curves of section II.2.

B. I_s and I_{diff} Neglecting Nonradiative Recombinations

As shown in figure (6) for the optical cavity region beyond the strip contact ($x \geq w$), there are two components to the lateral carrier flow. The first component is an ohmic sheet-current density I_s (A/cm), which flows in a thickness h of the p -layer and is driven by the p -layer voltage

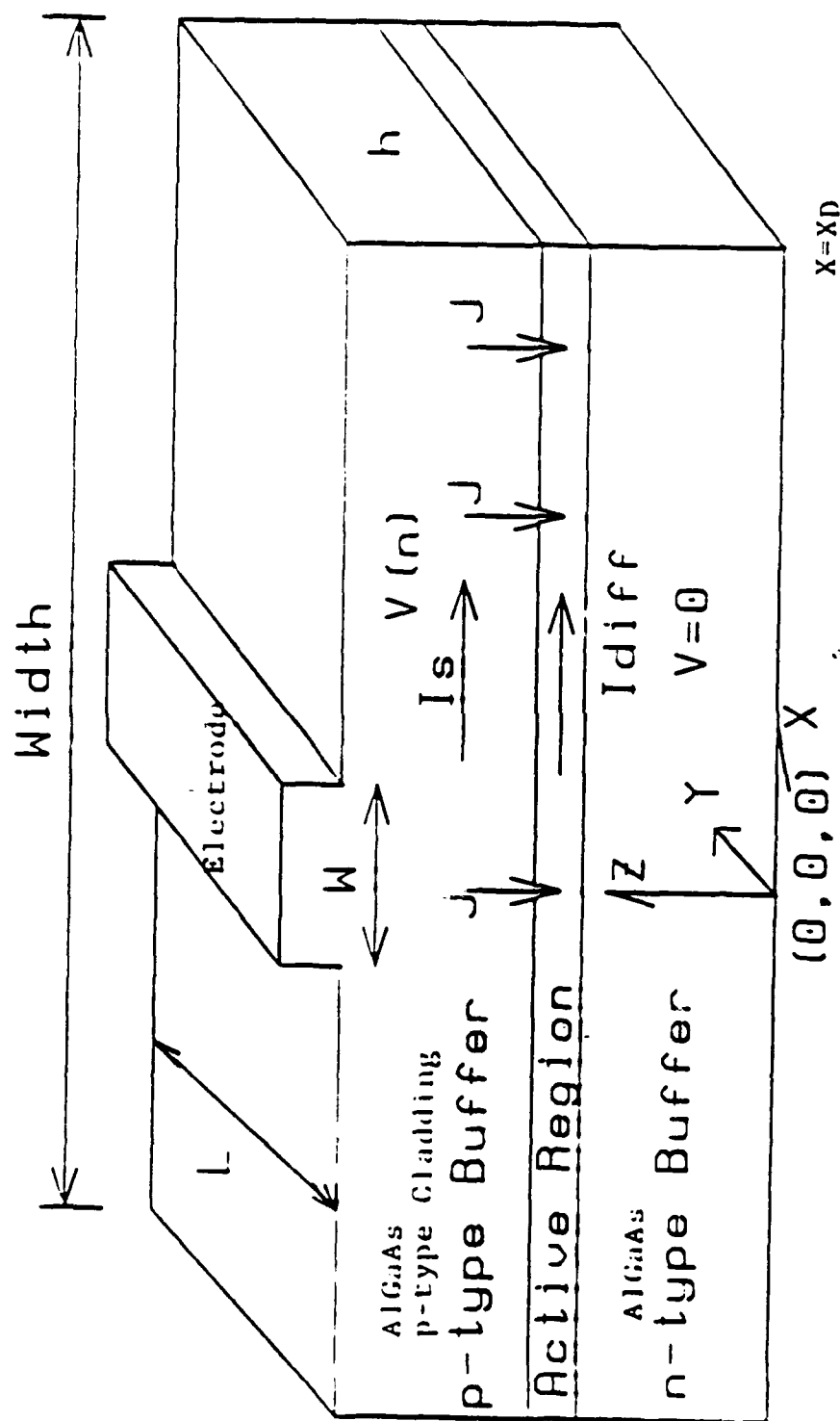


Figure (6): Current flow in a double-heterostructure strip-geometry laser. I_{diff} (A/cm^2) is the active region diffusion current density. I_s (A/cm) is the ohmic p-layer sheet-current density, and J (A/cm^2) is the injected current density.

gradient dV/dx , i.e.

$$\frac{dV(x)}{dx} = -\Omega I_s, \quad (21)$$

where Ω (ohms-cm) is the p-layer resistivity and ρ (ohms/square) is the sheet resistance. The sheet current density I_s leaves the p-layer becoming the active-region injected-current density $J(A/cm^2)$, i.e.,

$$\frac{dI_s}{dx} = -J. \quad (22)$$

It is assumed that the evolution of the electron concentration $n(cm^{-3})$ in the active layer can be described by the diffusion-type equation:

$$qD \frac{dn}{dx} = -I_{diff}, \quad (23)$$

where the effective diffusion coefficient $D(cm^2/s)$ may be either a constant or the concentration-dependent variable $D(n)$. To preserve charge neutrality the hole-diffusion-current density $I_{diff}(A/cm^2)$ and the electron-diffusion-current density $-I_{diff}(A/cm^2)$ sum to zero lateral current flow in the active region. The diffusion current has as its source the injection-current density J and, as its sink, the concentration-dependent recombination rate $R(n)(cm^{-3}/s)$; that is,

$$\frac{dI_{diff}}{dx} = \frac{J}{d} - q \left\{ R(n) + \frac{Pg[n(x)]}{\hbar\omega} \right\} \quad (24)$$

where q is the magnitude of the electron charge and d is the active-layer thickness. As a first order approximation any effect of the typically

smaller n -layer resistivity is neglected; thus the x -dependent p -layer voltage V is also the voltage across the active layer. The above equation must be solved subject to the self-consistent condition that, at each point x , the voltage V be a unique function, $V(n)$, of the local injected concentration n . Explicit functions for $V(n)$ and $R(n)$ will be introduced at a later stage, but controversy still surrounds the choice of these functions which are determined in a given heterostructure by the active region doping, by the amount and type of the nonradiative recombination. The above equations hold true for $x \geq w$ (beyond the strip) but by alternatively regarding J as the known function, one can find the solution under the strip. The additional sink term:

$$q \left\{ R(n) + \frac{Pg[n(x)]}{\hbar\omega} \right\}$$

in the above model represents a mode-stimulated recombination effect as well. Thus this model is applicable up to threshold or, above threshold, to the regions beyond the mode saturation. However, neither loss of confinement (carrier escape into the ternary layers) nor inhomogeneous temperature effects are included in this model. Boundary conditions at $x = \infty$ include $V = I_s = J = I_{diff} = dn/dx = 0$, and, by definition $n = n_\infty$. Several authors[23-28] have sought a description of lateral current spreading in heterostructures starting from equations that are

various particular cases of the model presented here. In each case the carrier flow in the active layer was assumed to result only from diffusion, i.e., any field driven carrier motion is tacitly neglected. This neglect does not appear to be accurate always, although is often is, particularly for an n-type active layer. The carrier transport in the active layer can be adequately modeled with diffusion equations as explained earlier. Although this model is similar to the model of Joyce [25], the following modifications have been introduced in our calculation.

- (1) Stimulated recombination is included as:

$$\left\{ R(n) + \frac{Pg[n(x)]}{\hbar\omega} \right\}$$

where P represents optical power density; $g[n(x)]$ is the gain expression for the carrier density n at position x calculated as before for QW structures; $\hbar\omega$ is the photon energy.

- (2) The specific voltage expression ($V(n)$) in QW structures is used:

$$V(n) = (E_{fc} + E_{fv} + E_{go})/q \quad (25)$$

with E_{fc}, E_{fv} calculated as shown in section II.2.

For narrow stripes, ($W = 4\mu m$ in our calculations), the assumption of a constant injected current density $\bar{J} = J_0$ under the strip can be adopted. Then one has to solve the above equations for each region separately: under the strip and beyond the strip. Nevertheless, the analytical solution used by Joyce cannot be applied because of the introduction of the stimulated term in the above equation. Therefore the Runge-Kutta numerical technique shall be adopted to solve the coupled, nonlinear, nonhomogeneous, differential equations.

C. Nonradiative Recombination Current

The above models does not account for nonradiative recombination effects for the calculation of the threshold current for lasing action. However, for comparison with the experimental results, nonradiative recombinations must be included. Thus, equation (24) is rewritten as:

$$\frac{dI_{diff}}{dx} = \frac{J}{d} - q \left\{ R(n) + \frac{Pg[n(x)]}{\hbar\omega} + \frac{n}{\tau_{nr}} \right\} \quad (26)$$

where τ_{nr} is the nonradiative recombination lifetime. Once again the Runge-Kutta numerical technique could be used for solving the diffusion current equation mentioned earlier. Then, I_s and I_{diff} , including the nonradiative recombination, are obtained where the approximation is made and the confinement factor in the y direction is approximated

by Γ_y . $g[n(x)]$ is calculated in the above using the effective index method in the case of the ridge waveguide structure of figure (2), we calculated the $E(x)$ expression. The effective index variation [26] induced by the carrier injection is calculated for the modal gain calculation. It must be noted that the absorption in the region below transparency is included in the above equation, where the expressed gain is the net modal gain, (g_{mod}) in the optical cavity.

D. Auger Recombination Current

Auger recombination current (J_{auger}) is modeled as [27]:

$$J_{auger} = qML_z C_a np^2 \quad (27)$$

with $C_a = 2 \cdot 10^{-30} cm^6/s$ for the CHSH Auger process. For other Auger processes see appendix A. The optical cavity recombination current (J_{cav}) is presented as [28]:

$$J_{cav} = q(2L_g) B n_{cav} (p_{cav} + p_o) \quad (28)$$

with $B = 9 \cdot 10^{-11} cm^3/s$, L_g represents the optical cavity width, p_o the doping level of the cavity, and n_{cav}, p_{cav} , respectively the electron and hole density in the optical cavity evaluated in a 3-D configuration from the relative position of the Fermi level in the QWs.

II.4 Threshold Current

In this section a general expression of threshold current (I_{th}) for MQW structure will be established. This current is composed of:

- (1) a radiative recombination current $J_r(A/cm^2)$ so that the modal gain, (g_{mod}) overcomes the cavity losses,
- (2) a lateral spreading current in the upper confinement layer $I_s(A/cm)$,
- (3) a lateral diffusion current in the active layer $I_{diff}(A/cm^2)$,
- (4) a surface or interface nonradiative recombination current, $J_{nr}(A/cm^2)$,
- (5) an Auger recombination current $J_{auger}(A/cm^2)$, and
- (6) an optical cavity recombination current $J_{cav}(A/cm^2)$ due to radiative recombination carriers leaking in the optical cavity.

Hence, the threshold current is expected as:

$$I_{th} = \{J_r(\text{forg} = g_{mod}) + J_{nr} + J_{auger} + J_{cav}\}WL \\ + 2I_sL + 2I_{diff} \cdot ML_zL \quad (29a)$$

with

$$g_{mod} = \alpha_i + \frac{1}{2L} \cdot \ln\left(\frac{1}{R_1 R_2}\right) \quad (29b)$$

Neglecting the nonradiative recombination current, the threshold current in the structure is approximated as:

$$I_{th} = (J_o + J_{auger} + J_{cav})WL + 2I_sL \quad (29c)$$

where the lateral diffusion current I_{diff} is included in $J_o (= J_r + J)$. To calculate the threshold current, the following process is used. For a given injection carrier density at the strip edges (n_e):

- (1) J_o and I_s are calculated as described before.
- (2) J_{auger} , J_{cav} are calculated as given in steps (3-4) below.
- (3) L^* can be calculated by rewriting the threshold condition:

$$L^* = \ln(1/R_1 \cdot R_2) \cdot \left\{ 2(g_{mod} - \alpha_i) \right\}^{-1} \quad (30)$$

(4) The threshold current I_{th} is obtained as per equation (29).

The following parameter values are used in the calculation:

1. Optical power density: $P = 10^5 W/cm^2$;
2. Sheet resistivity: $\rho = 0.50 \Omega cm$, and
3. Diffusion coefficient in the active layer: $D = 10 cm^2/s$,
4. $R_1 = 0.3$ and $R_2 = 0.8$ (one facet is uncoted and other is with treatment),

finally

5. α_i represents cavity losses.

II.5 Calculation of Modal Gain (g_{mod})

In this study the modal gain is calculated by:

$$g_{mod} = \frac{\Gamma_y \int_{-\infty}^{\infty} g[n(x)] E^2(x) dx}{\int_{-\infty}^{\infty} E^2(x) dx} \quad (31)$$

where the approximation $g[n(x, y)] \simeq g[n(x)] \cdot g[n(y)]$ is made and the confinement factor in the y-direction is approximated by Γ_y . $g[n(x)]$ is calculated as in the earlier expression for $g(E)$. Using the effective index

method in the case of the ridge waveguide structure $E(x)$ is calculated. The effective index variation [16] induced by the carrier injection is estimated to be less than 0.01 percent and is consequently neglected for the modal gain calculation. It must be noted that the absorption in the region below transparency is included in the previous expression, where the expressed model gain is the net modal gain in the optical cavity.

II.6 Mode Confinement Factor for MQWs

One of the main difference between the SQW and MQW is that the confinement factor (Γ) of the optical mode is significantly smaller for the former than that for the latter. This can result in a higher threshold carrier density and in higher threshold current density of SQW lasers when compared with the MQW lasers. The confinement factor of the fundamental mode for a SQW is:

$$\Gamma \approx D^2 / (2 + D^2), \quad (32)$$

with

$$D = \left\{ k_o \left(\mu_a^2 - \mu_c^2 \right)^{1/2} \cdot d \right\}, \quad (33)$$

is found to be accurate to within 1.5% [29]. In the above expression d is the active layer thickness, $k_o = 2\pi/\lambda_o$ where λ_o is the wavelength in

free space and μ_a , μ_c are the refractive indices of the active and the cladding layer, respectively. For smaller active layer thickness $D \ll 1$ and Γ is expressed as[30]:

$$\Gamma \cong 2\pi^2(\mu_a^2 - \mu_c^2) \cdot d^2 / \lambda_o^2. \quad (34)$$

The mode confinement in MQW structures can be analyzed by solving the electromagnetic wave equation for each of the layers with the appropriate boundary conditions. This procedure is quite tedious and Streifer et al.[31] have shown that the following simple formula gives reasonably accurate results:

$$\Gamma(MQW) = \gamma \frac{N_a d_a}{N_a d_a + N_b d_b}, \quad (35)$$

where

$$\gamma = 2\pi^2(\bar{\mu}^2 - \mu_c^2) \left\{ N_a d_a + N_b d_b \right\}^2 / \lambda_o^2, \quad (36)$$

and

$$\bar{\mu} = \frac{N_a d_a \mu_a + N_b d_b \mu_b}{N_a d_a + N_b d_b}, \quad (37)$$

N_a , N_b are the number of active and barrier layers in the MQW structure and d_a , d_b (μ_a and μ_b) are the thickness (and refractive indices) of the active and barrier layers, respectively. $\bar{\mu}$ is the average refractive index of the uniform optical mode in the MQW active region. Hence the

γ is the confinement factor of the optical mode. Γ is thus obtained by multiplying γ by the ratio of the total active thickness to the total thickness of the active and barrier layers.

Chapter III

Calculation And Optimization

Figure (3) is used for chracterization of MQW stucture under inves-
tigation. At this stage simple calculations are performed to predict the
device performance using the models of chapter II. However, extensive
calculations are required essentially to identify the model limitation, if
any, and use realistic values for the parameters to compare the experi-
mental data with the theorical calculations. Furthermore, optimization
of (1) device performance, and (2) structural parameters for device ge-
ometry are beyond the scope of this investigation. Although a scheme
has been identified for optimization in general. Furthermore, using the
models of chapter II an extensive study can be conducted.

For our immediate interest in device characterization, an eigenvalue
problem (quantum confined carrier energy levels) is solved numerically
using the Newton-Rapson iterative technique. Furthermore, the QW
barrier height is so small (even in the conduction band discontinuty)
it does not permit more than one quantized energy level in each of
the conduction and valence bands of GaAs/AlGaAs hetero structure.

Therefore, the model calculations are limited to a single quantized energy level in the QWs. Therefore, the behavior of one of the most important functions, namely, the nonlinear gain ($g_{max}(n)$), is not apparent from these calculations.

Fourth order Simpson's rule is employed for numerical integration of the gain expression. Notice, for a given QW layer thickness, the gain spectrum is shown in figure (4). Figure (5) is basically the variation in maximum gain as a function of injected current. From this preliminary calculation, it is evident that an abrupt transition (from zero gain to a finite gain) occurs just above the band gap energy, exhibiting the lowest energy quantization effect. Furthermore, a little above the abrupt transition energy level, there is a sharp peak representing the maximum gain, g_{max} , with bell-shaped spectrum of the system in general. This is followed by gain harmonics represented by multiple peaks of relatively smaller magnitude. By changing the well size (QW layers thickness L_z) the gain spectrum reveals a slight shift in energy level for maximum gain. Calculation accuracy is essential to see the pronounce shift in its energy for maximum gain conditions. Although the linewidth of the gain spectrum is very narrow, but the overall profile of the peak looks

like a bell-shaped curve. It is also evident that maximum gain is three orders of magnitude larger than its baseline effect.

III.1 Threshold Current Calculation

For the calculation of threshold current one needs to solve the differential equations with the boundary conditions for the models described in chapter II. These differential equations are coupled nonlinear and nonhomogenous with multipoint boundary conditions. In general, this requires a technique similar to two point boundary value problems to calculate the threshold current. The Shooting method needs to be employed to solve the multipoint boundary value problem in conjunction with the Runge-Kutta technique to solve the differential equations.

As a first order model approximation, the injection current (source term) is used to establish the boundary value requirements for maximum free electron concentration in the optical cavity under the influence of ridge waveguide of the QW structure. Furthermore, on the outer boundaries of the optical cavity, intrinsic carrier concentration of GaAs is assumed as boundary conditions imposed on the system. These realistic assumptions help in establishing the carrier concentra-

tion profile in the active region to estimate the sheet current, I_s , and diffusion current, I_{diff} . In our investigation, a Gaussian distribution is assumed for the carrier concentration in the active region of optical cavity along x -direction with the boundary conditions as mentioned before. Once gain, it is estimated that the Auger and cavity recombination currents are expected to be no more than 20% of the radiative recombination current, J_r . However, its contribution will not be more than 2% of the total threshold current, i.e., threshold current due to radiative recombination (including the cavity losses) and nonradiative recombination effects (I_s, I_{diff}) .

III.2 Design Optimization

Neglecting the Auger and cavity recombination effects, just to investigate the structural parameters, we could approximate the expression for threshold current as:

$$I_{th} = \left\{ J_o(M) \cdot W + 2I_s \right\} \cdot L^* \quad (38)$$

However there exist a relationship for optimal cavity length, L^* as:

$$L^* = \frac{\ln(1/R_1 R_2)}{2(\Gamma(M)g_{max} - \alpha_i)}$$

where L^* represents an optimal optical cavity length. Since $\alpha_i \ll \Gamma(M) g_{max}$ and considering that:

$$J_o(M) = M J_{o-SQW} \quad (39)$$

we have

$$I_{th} = \frac{M}{\Gamma(M)} \left\{ J_{o-SQW} \cdot W + \frac{2I_s}{M} \right\} \cdot \frac{\ln(1/R_1 R_2)}{2\zeta_{max}} \quad (40)$$

J_{o-SQW} represents the injection current in n SQW layers. This expression shows that I_{th} is a nonlinear function of M , $\Gamma(M)$, and g_{max} . Furthermore, the net effect of I_s can be reduced significantly by increasing the number of QW layers. Fine tuning of this code is essential to exploit the power of this model.

Chapter IV

Conclusions

Analytical and computer models are developed to estimate the influence of each parameter for the calculation of the threshold current. However, realistic data of various parameters are required to fully understand their effects and design for optimal performance. These parameters can be summarized in the following:

- (1) confinement factor,
- (2) ridge structure parameters,
- (3) QW structure parameters,
- (4) role of leakage current components on the threshold current.

As an example, the confinement factor is a function of:

number of quantum well layers.

QW layer thickness of the active region.

X : Aluminum fraction in the barrier.

Y : Aluminum fraction in the buffer layers for gain confinement.

L_b : QW barrier width.

L_g : Optical cavity width.

W : strip width.

h : thickness of active material in the MQWs.

Regarding the ridge structures, one needs to investigate the parameters associated with the optical cavity and numbers of QWs as listed below.

(i) Optical Cavity Parameters

This is once again function of the following:

X : Aluminum fraction in the barrier.

Y : Aluminum fraction in the buffer layers for confinement effect.

L_b : QW barrier width.

L_g : Optical cavity width.

(ii) Influence of Well Number (M)

In this case, for each value of M one may find an optimal value of $L = L^*(M, L_z)$ corresponding to the minimum value of I_{th} . As a matter of fact the existence of this optimal value $M^*(L_z)$ for a given L_z is explained earlier in section II.2, "Design Optimization". However, investigation is needed for the threshold current as a function of cavity length ($I_{th}(L)$) for a given value of L_z . For parameter optimization of QWs see steps (3a) and (3b).

(3a) Optimization of Well Width L_z For Specific M

By repeating the calculation of step (2) for a different L_z , a set of $\{M^*(L_z), L_z\}$ pairs can be obtained.

(3b) Optimization of Al Fraction In Cladding Layers

This program permits for variation in Al fraction Y in the cladding (buffer) layers. However, different values of Y could lead to sensitive variations for the optimal threshold current.

Finally role of leakage current would shed some light on threshold current requirements.

(4) Leakage Current Effects

Leakage current components such as (I_{diff} , I_{diff} , I_{auger} , I_{cav}) can be investigated for various parameters listed earlier using this computer code.

(5) Direction For Further Investigations

One can extend this work to model the entire optical switch which consists of a main laser and two side lasers once the optimized geometry is established as shown earlier. Before this can be done some fine tuning of the computer model is essential, which requires structural parameters and experimental data to verify the predicted results.

Chapter V

Computer Program Listing

Standrad Fortran 77 language is used to develop this code. However, this code can work on Apollo 3500 or DNK 10000 machine, under Unix operating system. Initially this code was developed for IBM pc, which requires at least a day to perform the calculation for gain spectrum of figure (4). Therefore, this code is transported to Appollo 3500 and computational time was significantly reduced to 3 hours. However, about one hour is required to run this program on DNK 10000. Further testing of this software is required to make sure that all the bugs are out and program is efficient.

```

program lasergain
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mlh,lz,men,mehx,mx
character*1 aa(51)
dimension ee(4),hh(4),eh(4),egn(4),gte(200,4),
1 tt(2,512,512),xmax(2),xmin(2),af1(2),xxnn(25),xxpp(25),
2 aeven(2),aodd(2),sum(2),xxlz(10),gm(25,5),xjcav(25),xjaug(25),
3 v(25),xidif(25),xj(25),xis(25),xdum(25),wwlz(6),cj(25,2)
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mlh,er,ego,vo,voo,lz
common /b/meh,mehx,en1
common /c/mx,evj,p2
common /d/ee,hh,ieh
common /e/xme,xmh,hb1,hb12,pil,pil2,xkt,xlz,ec,ev,xnn,xpp
common /f/tt,kk,mm
common /g/gte,xmax,xmin,aa
common /input/xlength,xwidth,xw,wlz,wlb,buffer,buffer,ddd,
1 pdphot,rsheet,rr1,rr2,difc,alpi,cinj,xwd,xxnn,xxpp,xnpo,xnpi,
2 xno,xpo,cin1,cin2,a,nwell,nnx

```

This file is: aqw.test.3.1
Updated on 2/4/1992

```

open(5,file='//hep10k/user/khaja/aqw.test.3.in',status='old')
open(6,file='//hep10k/user/khaja/aqw.test.3.1.out',status='new')
1 )
open(7,file='//hep10k/user/khaja/aqw.test.3.1.plot.out',
1 status='new')
156789012345678901234567890123456789012345678901234567890123456789012
write(6,*)'
write(6,*)' Source-code file: aqw.test.3.1 '
write(6,*)' Output file: aqw.test.3.1.out '
write(6,*)'
x=0.d0
refa=ref(x)
x=0.2d0
refb=ref(x)
x=0.40d0
wlb=30.d-9
wlg=100.d-6
refw=ref(x)
wwlz(5)=300.d-9
wwlz(4)=250.0d-9
wwlz(3)=200.d-9
wwlz(2)=150.d-9
wwlz(1)=100.d-9
voo=0.12d0
vo=voo
t=300
dj=5.d-5
ic=0
nwell=4
ijx=1
injp=25
injp must be 25
nnx=25
xnpi=2.1d12
xnpo=1.d6/(50.d0*400.d0*1.6029d-19)
nsize=5

```

```

inwl=3
buffp=200.d-9
do 900 nws=1,nwsiz
lz=wwlz(nws)
wlz=lz
xlz=lz
do 855 inwell=1,inwl
nwell=inwell*2
ddd=dfloat(nwell)*lz
cinj=0.d0
lzp=lz*1.d9
itest=0
write(6,*)' Lz=',lzp,'(nm), nwell=',nwell
write(6,*)' Index: (GaAs)=' ,refa,' , (AlGaA)=' ,refb,
1  ' , (p)=' ,refw
write(6,*)' I(inj) I(th) <Idif> <Is> <J> ',
1  ' <gmod> <Iaug> <Icav>'
1456789012345678901234567890123456789012345678901234567890123456789012
do 850 inj=ijx,injp
itest=itest+1
cinj=cinj+dj
call input
ic=ic+1
if(itest.gt.1)go to 25
call elece(ic)
thicn=lz*dfloat(nwell)+wlb*dfloat(nwell-1)+buffp
refave=(lz*refa+wlb*refb+buffp*refw)/thicn
xnbar=1.d0
25 continue
do 800 inn=1,nnx
xnn=xxnn(inn)
xpp=xnn
x=.2
xme=fme(x)
xmh=fhh(x)
hb1=hb
hb12=hb1*hb1
pil=pi
pil2=pil*pil
taw=1.d-12
xht=hb/taw
xht2=xht*xht
evj2=evj*evj
evj12=dsqrt(evj)
xconst=3d0*hb*q*q/(xme*xme*c*ego*xnbar)
xkt=k*t
xmo=1.333*mo*ego*evj
xrm=xme*xmh/(xme+xmh)
n=0
xmax(1)=-1.d10
xmin(1)=1.d10
xmax(2)=-1.d10
xmin(2)=1.d10
50 n=n+1
ecn=ee(n)
ec=ego
evn=hh(n)
ev=0.d0
egn(n)=ecn+evn+ego
egnn=egn(n)
evj1=evj
efc=a2inv(evj1)
efv=a4inv(evj1)

```

```

a=egnn
a1=a
b1=1.5d0*a
b2=100.d0*a
50 mm=2000
mm1=mm+1
dex2=(b2-a)/dfloat(mm1)
ijk=200
dex1=(b1-a1)/dfloat(ijk)
den1=ee(n)+hh(n)
xec=0.d0
xev=0.d0
do 200 i=1,ijk
  ephon=a1+dex1*dfloat(i-1)
  xp=.0d0
  afl(1)=0.d0
  afl(2)=0.d0
  aeven(1)=0.d0
  aodd(1)=0.d0
  aeven(2)=0.d0
  aodd(2)=0.d0
  do 100 j=1,mm1
    ecvx=egn(n)+dex2*dfloat(j-1)

    if(ecvx.gt.b1) go to 101

    x1=red(xrm,den1,ephon,egnn)*evj12
    eipcn=ecn+(ecvx-egn(n))*xrm/xme
    x21=xmte(xmo,ecn,eipcn)
    x22=xmtm(xmo,ecn,eipcn)
    x3=fc(xrm,xme,ecvx,egnn,xec,efc,xkt)
    x4=fv(xrm,xmh,ecvx,egnn,xev,efv,xkt)
    x5=(ecvx-ephon)*evj
    x6=x5*x5+xht2
    xte=x1*x21*(x3-x4)*xht/(x6*ecvx*evj)
    xtm=x1*x22*(x3-x4)*xht/(x6*ecvx*evj)
    if(j.eq.1)go to 68
    if(j.ne.mm1)go to 70
58  afl(1)=afl(1)+xte
    afl(2)=afl(2)+xtm
    go to 80
70  if(j.ne.j/2*2)go to 75
    aeven(1)=aeven(1)+xte
    aeven(2)=aeven(2)+xtm
    go to 80
75  aodd(1)=aodd(1)+xte
    aodd(2)=aodd(2)+xtm
10  continue
    if(xte.lt.xp)go to 100
    xp=xte
10  continue
11  continue
    do 102 j=1,2
      sum(j)=afl(j)+4.d0*aodd(j)+2.d0*aeven(j)
2   sum(j)=sum(j)*dex2*evj*xconst/3.d0
    xxx=sum(1)
    gte(i,1)=ephon
    gte(i,2)=sum(1)
    gte(i,3)=sum(2)
    if(xxx.le.xmax(1))go to 110
    xmax(1)=xxx
    imax1=i
    go to 115

```



```

110 if(xxx.ge.xmin(1))go to 115
    xmin(1)=xxx
115 continue
    xxx=sum(2)
    if(xxx.le.xmax(2))go to 120
    xmax(2)=xxx
    imax2=i
    go to 200
120 if(xxx.ge.xmin(2))go to 200
    xmin(2)=xxx
200 continue

    x1=dfloat(nwell)*w1z
    x2=dfloat(nwell-1)*w1b
    x3=x1+x2
    avmu=(x1*refa+x2*refb)/x3
    gama=2.d0*pi*pi*(x3**2)*(avmu**2-refw**2)
    xlamb=hb*c/(gte(imax1,1)*evj)
    gama=gama/xlamb/xlamb
    cf=gama*x1/x3/avmu
    x11=w1z
    xv=vo
    if(inn.gt.0)go to 430
    write(6,*)' '
    write(6,*)'          *** gain spectrum *** '
    write(6,*)' '
    write(6,*)'      Lz=',x11,' [m],  Vo=',xv,' [eV]'
    write(6,*)'      n =',xnn,' [m3], p =',xpp,' [m3]'
    write(6,*)'      T =',t,' [oK]',',', ' = well',nwell
    do 400 i=1,ijk
    gte(i,2)=gte(i,2)*cf
    gte(i,3)=gte(i,3)*cf
    if(inn.gt.1)go to 400
    write(6,*)gte(i,1),gte(i,2),gte(i,3)
400 continue
    ktem=1
    imax=imax1
420 call plot(ijk,ktem,imax)
    write(6,*)' '
    ktem=2
    imax=imax2
    call plot(ijk,ktem,imax)
    write(6,*)' '
430 continue
    xpp=xxpp(inn)
    efv=a4inv(evj1)
    v(inn)=efc+dabs(efv)+ego
    cj(inj,1)=cin1*1000.d0
    cj(inj,2)=cin2*1000.d0
    gm(inn,1)=gte(imax1,1)
    gm(inn,2)=gte(imax1,2)
    gm(inn,3)=gte(imax1,3)
300 continue
234567890123456789012345678901234567890123456789012345678901234567890123456789012
x=0.d0
xidif(1)=0.d0
xis(1)=0.d0
xx1=gm(2,1)**2+gm(3,1)**2
xx=dsqrt(xx1)
xj(1)=q*(pdphot*xx/(gm(1,1)*evj)+(xxnn(1)-xnpi)/tau)*ddd
do 810 i=2,nnx
x=x+xxd
xx1=gm(i,2)**2+gm(i,3)**2

```

```

xx=dsqrt(xx1)
spn=pdphot*xx/(gm(i,1)*evj)
xrn=(xxnn(1)-xnpi)/tau
parg=2.d0*(a/xw)**2
arg=parg*x*x
xnxo=xno*dexp(-arg)
xnx=xnxo+xnpi
dxnx=xnx*(-parg*x)
d2xnx=xnx*parg*x*(parg*x-1.d0)
xj(i)=q*ddd*(dife*d2xnx+xrn+spn)
xidif(i)=-q*dife*dxnx
0 continue
xis(1)=0.d0
nnx1=nnx-1
do 812 i=2,nnx1
x6=(v(i+1)-v(i-1))/(2.d0*xwd*rsheet)
2 xis(i)=-x6
i=nnx
xis(i)=-(v(i)-v(i-1))/(xwd*rsheet)
do 813 i=1,nnx
xjaug(i)=xxnn(i)*xno*xno*(2.d-42)*q*dfloat(mwell)*lz
3 xjcav(i)=q*wlg*2.d0*(9.d-17)*(xxpp(i)+xnpo)*xnpi*xnpi
1 /xxpp(i)*9.d-17
xxl=xwidth*.5d0
call tsum(xidif,xwd,xxl,val)
avedif=val
va2=val*1.d-4
xxl=xwidth*0.5d0
call tsum(xj,xwd,xxl,val)
avexj=val
xxl=xwidth*.5d0
call tsum(xis,xwd,xxl,val)
avexis=val
xitha=(avexj*xwidth+2.d0*avedif*ddd)*1.2d0*xlength
xithc=2.d0*avexis*xlength
xith=xitha+xithc
do 830 i=1,nnx
xx=gm(i,2)**2+gm(i,3)**2
xx=dsqrt(xx)
0 xdum(i)=xx*xxnn(i)
xxl=xw
call tsum(xdum,xwd,xxl,val)
aveg=val*nwell
do 840 i=1,nnx
0 xdum(i)=xxnn(i)
call tsum(xdum,xw,xxl,val)
avenn=val
gmod=aveg/avenn
xxl=xwidth*0.5d0
call tsum(xjaug,xwd,xxl,val)
avjaug=val
call tsum(xjcav,xwd,xxl,val)
avjcav=val
cjcav=avjcav*wlg*xwidth
cjaug=avjaug*wlg*dfloat(mwell)*xlength*xwidth
rr=rr1*rr2
rr=1.d0/rr
oplen=2.d0*dlog(rr)/gmod
cin=cinj*1000.d0
cth=xith*1000.d0
cdif=avedif*1.d-4
cis=avexis*.01d0
cjd=avexj*1.d-4

```

```

caug=cjaug*1.d9
ccav=cjcav*1.d9
write(6,842)cin,cth,cdif,cis,cjd,gmod,caug,ccav
142 format(2x,f5.2,3g11.4,2f6.3)
150 continue
155 continue
160 continue
165 close(5)
close(6)
close(7)
stop 555
end
subroutine tsum(xx,xdx,xl,val)
implicit real*8(a-h,o-z), integer*4(i-n)
dimension xx(25)
xf=xx(1)-xx(25)
xodd=0.d0
xeven=0.d0
do 20 i=2,24
if(i.eq.i/2*2)go to 10
xodd=xodd+xx(i)
go to 20
10 xeven=xeven+xx(i)
20 continue
val=(xf+4.d0*xodd+2.d0*xeven)*xdx/3.d0
val=val/xl
return
end
subroutine fill(ii)
implicit real*8(a-h,o-z), integer*4(i-n)
character*1 aa
dimension gte(4,200),xmax(2),xmin(2),aa(51)
common /g/gte,xmax,xmin,aa
do 20 i=1,51
20 aa(i)=' '
aa(1)=':'
do 30 i=2,ii
30 aa(i)='*'
return
end
subroutine input
implicit real*8(a-h,o-z), integer*4(i-n)
dimension xxnn(25),xxpp(25)
common /input/xlength,xwidth,xw,wlz,wlb,bufp,bufn,ddd,
1 pdphot,rsheet,rr1,rr2,difc,alpi,cinj,xwd,xxnn,xxpp,xnpo,xnpi,
2 xno,xpo,cin1,cin2,a,nwell,nnx
456789012345678901234567890123456789012345678901234567890123456789012

```

mks units are used

```

xlength=200.d-6
xwidth=80.d-6
xw=xwidth*0.05d0
wlb=50d-9
pdphot=10.d5
rsheet=50.d0
rr1=.85
rr2=.3
difc=1.d-3
alpi=300.d0
taw=1.d-12

```

```

q=1.60219d-19
a=0.2d0
xnp1 = QW intrinsic carrier density [per m cube]
xnp0 = p-buffer zone doping density [per m cube]
tau = carrier life time [s]
q = electronic charge [Q]
cinj = injected current for lasing action [A]
xlength= laser cavity length [m]
xw = width [m]
xwidth = total width laser chip [m]
nwell = number of qws
wlz = qw width [m]
wlb = barrier width [m]
buffp = p-type buffer thickness [m]
buffn = n-type [m]
pdphot = optical power density [watts/m square]
rsheet = p-type buffer sheet resistance [ohms-m]
rr1 & rr2 = mirror reflectivity
dffc = diffusin constant [m square/s]
alpi = cavity losses [per m]

```

```

cinl=cinj
xj=cinl/(xw*xlength)
ad=ddd
xno=tau*xj/(q*ad)
xwd=xw*0.5d0
xn=xwidth/xwd
nnx=int(xn)+1
if(nnx.gt.25)nnx=25
xwd=xwidth/dfloat(nnx)
x=0.d0
xw2=xw*.5d0
do 20 i=1,nnx
x=x+xwd
if(x.gt.xw2)go to 10
xxnn(i)=xnp1+xno
xxpp(i)=xnp0-xno
if(xpo.gt.xnp0)xxpp(i)=1.d0
go to 20
xx=a*x/xw
xx=xx**2
x2=xno*dexp(-xx)
xxnn(i)=xnp1+x2
xxpp(i)=xnp0-x2
if(xpo.gt.xnp0)xxpp(i)=1.d0
continue
do 30 i=1,nnx,3
j1=i
j2=j1+2
if(j2.gt.nnx)j2=nnx
write(6,*)(xxnn(j),j=j1,j2)
write(6,*)' hole density [ m inverse cube] '
do 33 i=1,nnx,3
j1=i
j2=j1+2
if(j2.gt.nnx)j2=nnx
write(6,*)(xxpp(j),j=j1,j2)
return
end
subroutine plot(ijk,ktem,imax)
implicit real*8(a-h,o-z), integer*4(i-n) 6-47
dimension gte(200,4),aa(51),xmax(2),xmin(2)
character*1 aa

```

```

common /g/gte,xmax,xmin,aa
if(xmax(ktem))5,600,5
5 if(ktem.eq.2)go to 10
write(6,*)'
write(6,*)'   ***   qw gain vs photon energy plot   ***'
write(6,*)'   ***   transfer electric spectrum   ***'
write(6,*)'           Max=',xmax(1),', Indx=',imax
write(6,*)'           Min=',xmin(1)
write(6,*)'
ik=ktem+1
go to 20
0 write(6,*)'
write(6,*)'   ***   qw gain vs photon energy plot   ***'
write(6,*)'   ***   transfer magnetic spectrum   ***'
write(6,*)'           Max=',xmax(2),', Indx=',imax
write(6,*)'           Min=',xmin(2)
write(6,*)'
ik=ktem+1
0 continue
ial=imax-25
if(ial.lt.1)ial=1
ia2=imax+25
if(ia2.gt.ijk)ia2=ijk
xx=xmax(ktem)
56789012345678901234567890123456789012345678901234567890123456789012
write(6,*)'   E(ph)           g(E)           *****   Plot   *****'
write(6,*)'
do 500 is=ial,ia2
xe=gte(is,ik)
xd=xe/xx*50.
ix=int(xd)+1
if(xe.lt.xx2)go to 500
if(ix.gt.51)ix=51
call fill(ix)
write(6,505)gte(is,1),xe,aa
0 continue
5 format(2x,f7.3,g12.4,2x,51a1)
return
0 write(6,605)xmax(ktem)
5 format(2x,' ** zero maximum gain (' ,g20.5,')   **')
write(6,*)'
return
end
real*8 function red(xrm,d,e,egn)
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mih,lz,meh,mehx,mx
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mih,er,ego,vo,voo,lz
pi2=pi*pi
hb2=hb*hb
x=e-egn
if(x)20,10,10
0 x1=x/d
x2=dsqrt(x1)
d2=dsqrt(d)
x3=xrm*2.d0/hb2
x32=dsqrt(x3)
x33=x32*x32*x32
xx=x33*d2*int(x2)/(pi2*2.d0)
red=xx
return
0 write(6,*)' e=',e,' and egn=',egn
write(6,*)' x=',x,', x1 < 0 in red routine'
stop 100

```

```

end
real*8 function xmtm(xmo,ecn1,eipcn)
implicit real*8(a-h,o-z), integer*4(i-n)
common /e/xme,xmh,hb1,hb12,pil,pil2,xkt,xlz,ecn,ecv,xnn,xpp
x=xmo*(1.d0-ecn1/eipcn)*1.5d0
xmtm=x
return
end
real*8 function xmte(xmo,ecn1,eipcn)
implicit real*8(a-h,o-z), integer*4(i-n)
common /e/xme,xmh,hb1,hb12,pil,pil2,xkt,xlz,ecn,ecv,xnn,xpp
x=xmo*(1.d0+ecn1/eipcn)*.75d0
xmte=x
return
end
real*8 function a2inv(evj)
implicit real*8(a-h,o-z)
common /e/xme,xmh,hb1,hb12,pil,pil2,xkt,xlz,ecn,ecv,xnn,xpp
x=xnn*pil*hb1/xme*hb1/evj/xkt*xlz
if(x.lt.0.1d0)go to 10
y=dexp(x)-1.d0
efc=ecn+xkt*dlog(y)
a2inv=efc
return
y=x
j=1
do 20 jj=2,20
j=j*jj
x=x*x
y1=y
y=y+x/dbl(j)
dy=y-y1
if(dy.le.y*1.d-3)go to 25
continue
efc=ecn+xkt*dlog(y)
go to 5
end
real*8 function a4inv(evj)
implicit real*8(a-h,o-z), integer*4(i-n)
common /e/xme,xmh,hb1,hb12,pil,pil2,xkt,xlz,ecn,ecv,xnn,xpp
x=xpp*pil*hb1/xmh*hb1/evj/xkt*xlz
if(x.lt.0.1d0)go to 10
y=dexp(x)-1.d0
efv=ecv+xkt*dlog(y)
a4inv=efv
return
y=x
j=1
do 20 jj=1,20
j=j*jj
x=x*x
y1=y
y=y+x/dbl(j)
dy=y-y1
if(dy.le.y*1.d-3)go to 25
continue
efv=ecv+xkt*dlog(y)
go to 5
end
real*8 function xinte(ix,it)
implicit real*8(a-h,o-z), integer*4(i-n) 649
dimension tt(2,512,512)
common /f/tt,kk.mmm

```

```

      i=ix
      if(i.ne.1)go to 7
      do 5 j=1,mm,5
5      write(6,*)' tt(',j,', 1 )=',tt(i+1,j,1)
7      continue
      do 20 j=2,mm
      jx=4** (j-1)
      fjx=dfloat(jx)
      jxx=jx-1
      fjxx=dfloat(jxx)
      j1=j-1
      jj=j
      do 10 k=jj,mm
      k1=k+1
      tt(i,jj,k)=(fjx*tt(i,j1,k1)-tt(i,j1,k))/fjxx
10      continue
20      continue
      if(ix-1)50,30,50
30      ial=it-10
      if(ial.le.0)ial=1
      ia2=it+10
      if(ia2.gt.mm)ia2=mm
      write(6,*)' ** integrated values **'
      do 40 j=ial,ia2,3
40      write(6,*)' xinte: ',tt(i,j,j),tt(i,j+1,j+1),
1      tt(i,j+2,j+2)
50      continue
      xinte=tt(i,mm,mm)
      return
      end
      real*8 function dfloat(i)
      implicit real*8(a-h,o-z), integer*4(i-n)
      ix=i
      dfloat=dbl(i)
      return
      end
      real*8 function fc(xrm,xme,ecvx,egnn,ec,efc,xkt)
      implicit real*8(a-h,o-z), integer*4(i-n)
      x=xrm/xme*(ecvx-egnn)/xkt
      y=(efc-ec)/xkt
      if(x.gt.1.d5)go to 10
      x1=dexp(x-y)
      x2=1.d0/(1.d0+x1)
5      fc=x2
      return
10      x2=1.d-5
      go to 5
      end
      real*8 function fv(xrm,xmh,ecvx,egnn,ev,efv,xkt)
      implicit real*8(a-h,o-z), integer*4(i-n)
      xx=ecvx-egnn
      x=xx*xrm/xmh/xkt
      yy=efv-ev
      y=yy/xkt
      xx=-x+y
      if(xx.gt.1.d5)go to 10
      x1=dexp(-x+y)
      x2=1.d0/(1.d0+x1)
5      fv=x2
      return
10      x2=1.d-5
      go to 5
      end

```

```

subroutine eiee(ic)
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mlh,lz,meh,menx,mx
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mlh,er,ego,vo,voo,lz
common /b/meh,menx,enl
common /c/mx,evj,p2
dimension ee(4),hh(4),eh(4)
common /d/ee,hh,il

```

This part is from file qwee
as of 1/4/1992

23456

```

if(ic.gt.1)go to 10
pi=3.141593d0
p2=pi+pi
c=2.997925d8
epo=8.8542d-12
uo=1.26d-6
h=6.6262e-34
hb=1.05459e-34
q=1.60219d-19
mo=9.1095d-31
evj=1.60219d-19
mx=mo
k=8.62d-5

```

units: c[m], epo[f/m], uo[h/m], h[J.s], hb[J.s]
q[C], mo[kg], evj[conversion from ev to J],
k[eV/K degrees]

```

iee=0
ieo=0
inum=1
me=.0665*mo
mhh=.34*mo
mlh=.094*mo
er=12.9
ego=1.424
vvo=0.56
i1=2
10 i2=i1
vo=voo*evj
xx=0.
meh=fme(xx)
xx=.2
mehx=fme(xx)
vv1=vvo
in=1
if(ic.eq.1)go to 15
do 14 j=1,i1
14 eh(j)=ee(j)*.85*evj
15 continue
call energy(vv1,i2,eh,in,iee,ieo,inum,ic)
if(ic.eq.1)il=iee+ieo
do 20 i=1,i1
20 ee(i)=eh(i)/evj
xx=0.d0
meh=fhh(xx)
xx=.2d0
mehx=fhh(xx)
i2=i1
vv2=i.d0-vvo

```



```

in=i2
if(ic.gt.1)go to 22
ihe=1ee
iho=1eo
22 inum=2
if(ic.eq.1)go to 25
do 24 j=1,il
24 eh(j)=hh(j)*.85*evj
25 continue
call energy(vv2,i2,eh,in,ihe,iho,inum,ic)
do 30 i=1,il
30 hh(i)=eh(i)/evj
if(ic.gt.0)return
do 80 i=1,il
80 write(6,*)
1 '      E(' ,i ,', e) = ',ee(i)
do 90 i=1,il,2
90 write(6,*)
1 '      E(' ,i ,', h) = ',hh(i)
return
end
subroutine energy(vvo,ii,eh,in,ie,io,inum,ic)
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mlh,lz,meh,mehx,mx
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mlh,er,ego,vo,voo,lz
common /b/meh,mehx,enl
common /c/mx,evj,p2
dimension eh(4)
xp2=((pi*h/lz)**2)*0.5/meh
test=0.1d0
indx=1
2 continue

```

Calculations: electron/hole energy for odd number

```

do 10 i=1,ii,2
en1=((dble(i-1)*pi*h/lz)**2)*0.5/meh
en2=((dble(i)*pi*h/lz)**2)*0.5/meh
de=en2-en1
en1=en1+de*0.85d0
en2=en1
if(ic.gt.1)en2=eh(1)
vm=vvo*vo
ix=0
itest=0
5 ix=ix+1
if(ix.gt.999)go to 22
en1=en2
call qn1(vvo,v1,v2,itest)
if(itest.ne.0)go to 110
if(v2)6,27,6
6 den=-v1/v2
en2=en1+den
if(en2.lt.vm)go to 101
en2=vm*.95d0
101 continue
if(ix.le.2)go to 5
d2=dabs(en1*test)
if(dabs(den).gt.d2)go to 5
eh(i)=en2
go to(7,10),inum
7 if(ic.gt.1)go to 10
io=io+1

```

```

go to 10
10 eh(i)=en1*.95d0
go to 27
10 continue
write(6,*)' ix=',ix,' and eh(i)=',en2,' for odd i'
if(io.gt.1)go to 12
de=vm-en2

Test for difference in barrier height and electron
energy level computed just now

if(de.le.vm*0.1)return
12 continue
indx=1
go to (14,13),inum
13 if(ie.eq.0)return
14 continue
Calculations: electron/hole energy for even number

do 20 i=2,ii,2
en1=((float(i-1)*pi)*h/lz)**2)*0.5/mehx
enu=en1+xp2
en2=en1+xp2*.85d0
if(ic.gt.1)en2=eh(2)
ix=0
itest=0
15 ix=ix+1
if(ix.ge.999)go to 30
en1=en2
call qn2(vvo,v3,v4,itest)
if(itest.ne.0)go to 120
if(v4)16,30,16
16 den=-v3/v4
en2=en1+den
if(en2.lt.vm)go to 102
en2=vm*.95d0
22 continue
if(ix.le.2)go to 15
d2=dabs(en1*test)
if(dabs(den).gt.d2)go to 15
eh(i)=en2
go to(17,20),inum
17 if(ic.gt.1)go to 20
ie=ie+1
go to 20
20 eh(i)=en1*.95
go to 32
20 continue
write(6,*)' ix=',ix,' and eh(i)=',en2,', even i'
return
22 continue
write(6,*)' Out of loop 10 '
go to(26,27),indx
go to 27
26 indx=2
go to 2
27 il=1
eh(il)=en2
if(inum.eq.1)io=io+1
return
30 go to(31,32),indx
go to 32
31 indx=2

```

```

go to 14
12 i2=1
   eh(i2)=en2
   if(inum.eq.1) i2=i2+1
   return
end
subroutine qn1(vvo,f,df,itest)
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mlh,lz,meh,mehx,mx
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mlh,er,ego,vo,voo,lz
common /b/meh,mehx,enl
common /c/mx,evj,p2
ar=2.d0*meh*enl/h/h
a=dsqrt(ar)
ar=meh/enl*.5d0/h/h
da=dsqrt(ar)
vx=vvo*vo-enl
if(vx)20,20,10
10 ar=2.d0*mehx*vx/h/h
   b=dsqrt(ar)
   ar=mehx/2./vx/h/h
   db=-dsqrt(ar)
   ar=a*lz*0.5d0
   a1=dsin(ar)
   a2=dcos(ar)
   a3=a2/a1
   f=a*a3+b
   df=da*(a3-ar/a1/a1)+db
   itest=0
   return
20 vx=enl
   write(6,*)' enl=',enl,', vvo=',vvo,', and vo=',vo
   write(6,*)' vx=',vx,' < 0 in: qn1'
   itest=itest+1
   return
end
subroutine qn2(vvo,f,df,itest)
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mlh,lz,meh,mehx,mx
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mlh,er,ego,vo,voo,lz
common /b/meh,mehx,enl
common /c/mx,evj,p2
ar=2.d0*meh/h*enl/h
a=dsqrt(ar)
ar=meh/enl*.5d0/h/h
da=dsqrt(ar)
vx=vvo*vo-enl
if(vx)20,20,10
10 ar=2.d0*mehx*vx/h/h
   b=dsqrt(ar)
   ar=mehx/2./vx/h/h
   db=-dsqrt(ar)
   ar=a*lz*0.5d0
   a1=dsin(ar)
   a2=dcos(ar)
   a3=a1/a2
   f=a*a3-b
   df=da*(a3+ar/a2/a2)-db
   itest=0
   return
20 vx=enl
   write(6,*)' enl=',enl,', vvo=',vvo,', and vo=',vo
   write(6,*)' vx=',vx,' < 0 in: qn2'

```

```

itest=itest+1
return
end
real*8 function fme(x)
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mlh,lz,meh,mehx,mx
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mlh,er,ego,vo,voo,lz
common /b/meh,mehx,enl
common /c/mx,evj,p2
fme=(.0665d0+.083d05*x)*mo
return
end
real*8 function fhh(x)
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mlh,lz,meh,mehx,mx
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mlh,er,ego,vo,voo,lz
common /b/meh,mehx,enl
common /c/mx,evj,p2
fhh=(.034d0+.175d0*x)*mo
return
end
real*8 function flh(x)
implicit real*8(a-h,o-z), integer*4(i-n)
real*8 mo,k,me,mhh,mlh,lz,meh,mehx,mx
common /a/pi,c,epo,uo,h,hb,q,mo,k,me,mhh,mlh,er,ego,vo,voo,lz
common /b/meh,mehx,enl
common /c/mx,evj,p2
flh=(.094d0+.069d0*x)*mo
return
end
real*8 function ref(x)
implicit real*8(a-h,o-z), integer*4(i-n)
ref=3.590-0.170*x+0.091*x*x
return
end

```

List of References

- [1] W. Grande, "A study of GaAs/AlGaAs semiconductor laser devices for monolithic integrated optical circuits," Ph. D. dissertation, Cornell University, 1989.
- [2] W. T. Tsang, Appl. Phys. Lett., 39 (1981) 786.
- [3] O. Wada, T. Sanada, M. Kuno, and T. Fujii, Electron. Lett., 21 (1985) 1025.
- [4] R. Chin, N. Holonyak, Jr., B. A. Vojak, K. Hess, R. D. Dupuis, and P. D. Dapkus, Appl. Phys. Lett., 36 (1980) 19.
- [5] L. C. Chiu and A. Yariv, J. Luminescence, 30 (1985) 551.
- [6] N. Ogasawara, R. Ito, and R. Morita, Jpn. J. Appl. Phys. 24 (1985) L519.
- [7] Y. Arakawa and A. Yariv, IEEE J. Quantum Electron., QE-21, 1666 (1985).
- [8] R. Dingle, W. Weigmann and C. H. Henry, Phys. Rev. Lett. 33(1974) 827; R. Dingle and C. H. Henry, US Patent 3982207, September 21, 1976.
- [9] N. Holonyak Jr. R. M. Kolbas, R. D. Dapkus, IEEE J. Quantum Electron. QE-16(1980) 170.

- [10] N. Holonyak Jr, R. M. Kolbas, W. D. Laidig, B. A. Vojak, K. Hess, R. D. Dupuis and P. D. Dapkus, J. Appl. Phys. 51 (1980) 1328.
- [11] W. T. Tsang, Appl. Phys. Lett. 39 (1981) 786.
- [12] W. T. Tsang, IEEE J. Quantum Electron. QE-20 (1986) 1119.
- [13] S. D. Hersee, B. DeCremoux and J. P. Duchemin, Appl. Phys. Lett. 44 (1984) 476.
- [14] R. Dingle, Festkorperprobleme XV, H. J. Queisser, Ed. New York: Pergamon, 1975
- [15] R. Fivaz, J. Phys. Chem. Solids, 28 (1967) 839.
- [16] H. C. Casey, Jr. and M. G. Panish, Hetrostructure Lasers. New York: Academic, 1978.
- [17] H. Kroemer, Appl. Phys. Lett., 46 (1985) 504.
- [18] M. Asada, A. Kameyama, and Y. Suematsu, IEEE J. Quantum Electron., QE-20 (1984) 745.
- [19] N. K. Dutta, R. L. Hartman, and W. T. Tsang, IEEE J. Quantum Electron., QE-19 (1983) 1613.
- [20] M. Yamada and Y. Suematsu, Proc. 10th Conf. on Solid State Devices, Tokyo. 1978, Jpn. J. Appl. Phys. 18 (1979) Suppl.

18-1, 347.

- [21] M. Yamada, K. Hayano, H. Ishiguro and Y. Suematsu, Jpn J. Appl. Phys., 18 (1979) 1531.
- [22] M. Yamada and Y. Suematsu, IEEE J. Quantum Electron., QE-15 (1979) 743.
- [23] M. Yamada, H. Ishiguro and H. Nagato, Jpn. J. Appl. Phy. 49 (1978) 4644.
- [24] H. Kobayashi, H. Iwamura, T. Saku, and K. Otsuka, Electron Lett., 19 (1983) 166. and N. K. Dutta, R. L. Hartman, and W. T. Tsang, IEEE J. Quantum Elceton., QE-19 (1983) 1243
- [25] W. B. Joyce, J. Appl. Phys., 51 (1980) 2394.
- [26] See refrence 24
- [27] M. Takeshima, J. Appl. Physc., 58 (1985) 3846.
- [28] J. Nagle, Thesis presented at Paris VI University, October 1987.
- [29] D. Botez, IEEE J. Quantum Electron. QE-17 (1987) 178.
- [30] W. P. Dumke, IEEE J. Quantum Electron. QE-19 (1983) 932.
- [31] W. Streifer, D. R. Scifres and R. D. Burnham, Appl. Opt.

18(1979) 3547.

- [32] A. R. Beattie and P. T. Landsberg, Proc. R. Sco. London 249 (1959) 16.
- [33] W. T. Tsang, Appl. Physc. Lett. 38 (1981) 204.
- [34] N. K. Dutta and R. J. Nelson, J. Appl. Phys. 53(1982) 74.
- [35] R. J. Nelson and N. K. Dutta, in: Semiconductors and Semimetals, Vol. 22, Part C, ed. W. T. Tsang (Academic Press, New York, 1985) ch. 1.
- [36] [L. C. Chiu and A. Yariv, IEEE J. Quantum Elctron., QE-18 (1982) 1406..
- [37] N. K. Dutta, J. Appl. Phys., 54 (1983) 1236.

Appendix A
Nonradiative Recombination
Leakage Current-Auger Effect

An electron-hole pair can recombine nonradiatively meaning that the recombination can occur through any process that does not emit a photon. In many semiconductors, for example pure germanium or silicon, the nonradiative recombination dominates radiative recombination. The effect of nonradiative recombination on the performance of injection lasers is to increase the threshold current. If τ_{nr} is carrier lifetime associated with the nonradiative process, the increase in threshold current density is approximately given by:

$$J_{nr} = qn_{th}d/\tau_{nr}, \quad (A1)$$

where n_{th} is the carrier density at threshold, d is the active layer thickness and q is the electron charge.

A. Auger Effect

Since the pioneering work by Beattie and Landsberg [33], it is generally accepted that Auger recombination can be a major nonradiative mechanism in narrow-gap semiconductors. Recent attention to the Auger

effect has been in connection with the observed higher temperature dependence of threshold current of long-wavelength *InGaAsP* lasers compared to short wavelength *GaAs/AlGaAs* lasers. It is generally believed that the Auger effect plays a significant role in determining the observed higher temperature sensitivity of threshold current of *InGaAsP* lasers emitting near $1.3\mu m$ and $1.55\mu m$ [34,35]. In quantum well structures, the modification of the density of states may reduce the Auger recombination rate to allow long-wave-length QW lasers to have a low temperature sensitivity [36,37].

In the following, formulation for Auger rate calculation in QW is outlined. The various band-to-band Auger processes are shown in figure (A1). These processes are labeled CCCH, CHHS, and CHHL in which C stand for the conduction band, H for the heavy-hole band, L for the light-hole band and S for the spin split-off band. The energy versus wavevector (k_x, k_y) diagram is shown for one subband (one state in the z-direction) in each of the bands. The dashed curve (for CCCH) shows a second subband in the conduction band. The dashed line shows a process in which the electron makes a transition to a different subband. The CCCH process involves three electrons and a heavy hole and

is dominant in n-type material. The CHHS and the CHHL processes are dominant in p-type material. Under high-injection conditions as is present in lasers, all of the above mechanisms must be considered. However the Auger recombination rates for the CCCH, CHHS and CHHL processes have been previously calculated. In the nondegenerate approximation, the Auger rate R for the CCCH process varies as:

$$R \sim n^2 \cdot \exp(-\Delta E_c/k_B T), \quad (A2)$$

with

$$\Delta E_c = \frac{m_{ct} E_g}{m_v + 2m_c - m_{ct}}, \quad (A3)$$

where n, p are the electron and the hole concentrations respectively, m_{ct} is the effective masses of the electron at the higher energy E_t (electron 2' in figure (A1) and m_c, m_v are the band edge effective masses of the electron and hole, respectively. E_g is the energy different between the lowest conduction-band subband and the highest heavy-hole subband. For most cases, $E_t \cong E_g$ (band gap). The strongest temperature dependence and the band gap dependence of the Auger recombination rate is evident from the above equation. The Auger rate increases rapidly with increasing temperature and with decreasing band gap. For the CHHS

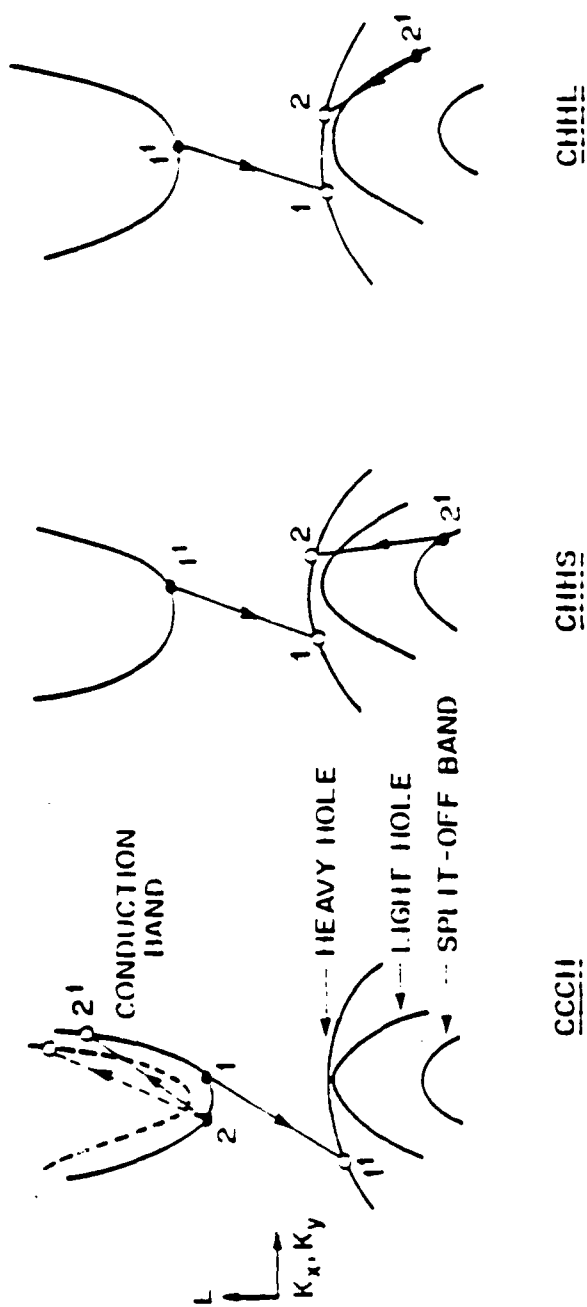


Fig. (A1): Band-to-band Auger recombination are shown schematically.

and CHHL processes, the Auger rate in the nondegenerate approximation varies as:

$$R \sim p^2 \cdot n \cdot \exp(-\Delta E_c/k_B T), \quad (A4)$$

with

$$\Delta E = \frac{m_s(E_g - \Delta_1)}{m_c + 2m_v - m_s}, \quad (A5)$$

for CHHS, and

$$\Delta E = \frac{m_l(E_g - \Delta_1)}{m_c + 2m_v - m_l}, \quad (A6)$$

for CHHL, where m_s, m_l are the effective masses of the split-off-band and light hole, respectively, and Δ_1 is the energy difference between the top of the heavy-hole band and that of the split-off band.

For undoped semiconductors, such as the active region of the semiconductor laser, the Auger rate at an injected carrier density n is given by:

$$R = C n^3, \quad (A7)$$

where C is known as the Auger coefficient. The Auger lifetime τ_A is given by:

$$\tau_A = n/R = \frac{1}{C n^2}. \quad (A8)$$

Thus the increase in the threshold current due to Auger recombination can now be obtained.

AFOSR RESEARCH INITIATION PROGRAM

FINAL REPORT

Fiber Optic Distribution System for Phased
Array Antennas

SUBMITTED TO: UNIVERSAL ENERGY SYSTEMS

SUBMITTED BY: DAVID A. SUMBERG
DEPARTMENT OF ELECTRICAL ENGINEERING
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NY 14623

SFRP RESEARCH LOCATION: RADC/DCLW, GRIFFISS AFB NY

USAF RESEARCHER: D. J. NICHOLSON

1. Abstract

A technique is described to equalize the optical path lengths for a multipath fiber optic distribution network. A novel path length adjuster is fabricated for a four channel system and used to provide continuous adjustment of the path lengths. Experimental data is presented.

2. Introduction

Phased array antennas play an important role in today's high performance communication and radar systems. Optical components, because of their light weight, wide bandwidth, and immunity to electromagnetic interference, are expected to play a major role in future phased array systems. This report addresses several problems that may be encountered in the implementation of a phased array with an optical distribution system.

One problem is the equalization of the microwave path lengths for the various channels of the system. A technique to measure the microwave path length is described in this report. The method uses a network analyzer to compare the phases of an RF modulated light signal in the different branches of the fiber optic distribution network. By equalizing the phases of all the branches, the path lengths may be equalized.

Another problem is that of providing a suitable broadband optical technique for adjusting the phase of each channel to an arbitrary value. One approach is to vary the length of the optical fiber by inserting a path length adjuster in the fiber path. This results in true-time-delay beamsteering. The path length adjuster described below consists of two GRIN lenses, one to collimate the beam and the other to refocus it into the fiber. It is shown to have low loss and can provide time delays of up to a nanosecond.

3. Theoretical and Experimental Design

3.1. True-Time-Delay Beam Steering System

A phase shift (β) may be introduced between successive antenna elements of a phased array by introducing a progressive time delay T_0 between each element. Such a time delay would result, for example, if the driving signal were split into N paths as in Fig. 3.1, each differing in length by an amount $L_0 = vT_0$, where v is the signal speed in the delay medium. If the frequency of the driving signal is f , the phase shift β is then given as

$$\beta = -2\pi T_0/T = -(2\pi T_0)f \quad (3.1)$$

where $T = 1/f$. The negative sign indicates that the phase of element $n+1$ lags that of element n .

3.2. Fiberoptic feed system for a phased array antenna

A block diagram for a four element phased array is presented in Fig. 3.2.

A signal source V_s modulates a laser diode (LD). The modulated optical signal is carried by a length of single-mode fiber (SMF) and divided into four paths of approximately equal length using three 1x2 single-mode couplers (C1, C2, C3). Each of the four beams is then collimated in free space using a graded index (GRIN) lens (G1-G4). The collimated beam traverses a variable distance L and is refocused by a second GRIN lens (G1'-G4') into a short length of multimode fiber that is pigtailed to a high speed photodetector (PD1-PD4). The subassembly consisting of the two GRIN lenses is referred to as a path length adjuster (PLA) and is described below. Finally, the signal is amplified (A1-A4) and delivered to the antenna elements (E1-E4).

3.3. Path Length Adjusters

The optical signal originates at the laser diode and is split into four fiber optic paths for delivery to the photodiodes. For a broadside beam each fiber optic path must be of equal length. Moreover, steering the beam requires a progressive increment in path length. Both of these requirements are satisfied by the path length adjuster (PLA) of Fig. 3.2. A more detailed drawing of the PLA can be found in Fig. 3.3.

The PLA consists of a collimated output beam from one leg of the coupler, a variable path length over which the beam travels, and a GRIN lens used to refocus the beam into the pigtail of a photodiode. The output of the coupler is collimated by a GRIN lens assembly that is especially designed for this experiment and is detailed in the next section. Both the output and input lens assemblies are mounted on a common optical axis using optical mounts which allow tilt and z (longitudinal) adjustment. In addition, the input lens is adjustable in the xy (transverse) plane.

Besides providing a means to steer the beam, the PLA also permits initial adjustment of the array for zero phase difference. Even though corresponding fibers in the four legs of the system are cut to nearly equal lengths, some uncertainty in length exists due to the variability within the couplers, the photodiodes, and the amplifiers.

It is possible to measure the length difference between any two legs of the array using a network analyzer. By measuring S_{12} , the length difference will manifest itself as a phase difference between legs. The PLA may be used to equalize the phases and thereby reduce the length differences to zero. Since the network

analyzer can resolve phase differences of less than 0.5° , at a frequency of 10 GHz it is possible, in theory, to resolve length differences as small as $40\text{ }\mu\text{m}$. In practice it is found that a length of $250\text{ }\mu\text{m}$ can be resolved with no difficulty, and with averaging this value may be lowered toward its theoretical limit.

3.4. GRIN Lens Collimators

A necessary condition for projecting the beam over a length of free space and refocusing it into a fiber is that the beam be well collimated. Also, for efficient coupling it is necessary that the core of the source fiber be no larger than that of the receiving fiber. The GRIN lens collimator of Fig. 3.4 was designed to meet these conditions.

The collimator consists of a glass sleeve, the elastomeric subsection of a GTE lab splice, and a 0.23 pitch GRIN lens. The fiber is centered in the glass sleeve and held firmly in place by the elastomeric splice. The fiber passes through the elastomeric splice and protrudes by about 0.5 mm. The GRIN lens is then positioned in the glass sleeve such that the emerging beam is well collimated. A more detailed assembly procedure may be found in section 4.3.2.

Two collimators, a source and a receiver, are required for each PLA. Single-mode fiber is used for the source collimator, because when used with a GRIN lens, single-mode fiber provides better collimation of the output beam than does multimode fiber. This follows since the HE_{11} mode is nearly Gaussian and has less beam divergence than that of multimode fiber. The mode field diameter for 850 nm single-mode fiber is less than $10\text{ }\mu\text{m}$, and this beam can be subsequently focused by a lens system into the multimode fiber of the receiver with very little loss.

The receiving collimator consists of a GRIN lens and a few meters of multimode fiber. Multimode fiber is used, because it provides higher coupling efficiency when used with the GRIN lens. The signal quality is unaffected by transmission over such short lengths of multimode fiber.

All GRIN lenses are anti-reflection coated for lower loss and to minimize back reflections into the laser diode.

4. Experimental results

4.1. Initial Path Length Adjustment

A procedure for equalizing all four path lengths using a network analyzer was described in section 3.3. One port of the analyzer drives the laser diode while the other port monitors the amplifier output. S_{12} is measured as the frequency is swept over an arbitrary range consistent with the bandwidth of the laser diode, photodiodes, and amplifiers.

Figure 4.1 shows the change in phase for one leg (channel 4) as the frequency is swept from 9 to 10 GHz. The markers record the phase at four specific frequencies for comparison with the remaining three channels.

Figures 4.2 - 4.4 show the phase change for the remaining three legs (channels 3, 2, and 1, respectively) with the phase change for channel 4 also appearing in each of the figures as an overlay. The markers are at the same frequencies as in Fig. 4.1. The figures were obtained after adjustments were made to equalize the lengths of channel 4, 3, 2, and 1 by bringing the phases of all four channels into coincidence.

The accuracy to which the path length difference between channels may be nulled depends on the accuracy to which the phase (β) may be determined. According to equation 3.1 the phase difference β may be written in terms of the length difference ΔL as

$$\beta = -2\pi(f/c)\Delta L \quad (4.1)$$

where $T_0 = \Delta L/c$ has been substituted. If β is expressed in degrees, this equation may be solved for ΔL as

$$\Delta L = - (\beta/360^\circ) \cdot (30/f) \quad (4.2)$$

where the frequency is in GHz and ΔL is in cm.

At an operating frequency of 10 GHz, eq. 4.2 indicates that a 1° variation in phase results in a path length difference of 0.008 cm or 80 μm . Figures 4.1 - 4.4 show an average phase variation of approximately 4° . Thus, the path lengths are equal to within approximately 300 μm .

4.2. True-Time-Delay Phase Shift

After equalizing the four path lengths as in section 4.1, the path length adjusters may be progressively incremented by an amount ΔL and the phase shift recorded and compared with theory. The value $\Delta L = 13.125 \text{ mm}$ is used, because it represents 60° beamsteering

when used with a microstrip antenna that has been designed for another phase of this project.

According to eq. 4.1, if the path length of each leg is incremented progressively by 13.125 mm, then the progressive phase delay (β) is given (in degrees) as

$$\beta = -15.75^\circ \cdot f \quad (4.3)$$

where f is expressed in GHz.

Thus β decreases linearly with f and has the value 141.75° at $f = 9$ GHz.

Figure 4.5 shows the measured phase shifts of channels 1-4 for frequencies 9.0-9.32 GHz. It is clear from the figure that the phase differences between channels remain constant for all frequencies, a required characteristic for a broadband system.

Table 4.1 lists the measured value of phase for each channel at 9 GHz and compares the progressive phase difference between adjacent channels with the theoretical value of 141.75° as found above.

Table 4.1

Measured phase of each channel and phase differences between adjacent channels at $f = 9$ GHz for $\Delta L = 13.125$ mm. Theoretical phase difference between adjacent channels is 141.75° .

channel	phase	measured phase difference
4	-119°	---
3	98°	143°
2	-45°	143°
1	176°	139°

Refer to Fig. 4.5 when computing the phase difference between channels. Note that phase decreases as we progress from channel 4 to channel 1. Starting at -119° at channel 4 proceed by 51° to -180° and then by 82° from -180° to $+98^\circ$ at channel 3. The phase shift from channel 4 to channel 3 is thus $51^\circ + 82^\circ = 143^\circ$. The remaining channels proceed in a similar manner.

According to Fig. 4.5 the phase decreases linearly with frequency (negative slope) as required by equation 4.1. The shorter positive slope segments of the curves are due to the discrete rather than continuous frequency sweep. Were the sweep to be continuous, then each negative slope segment would fall to -180° and rise sharply to $+180^\circ$ as frequency increased.

4.3. Grin Lens Collimator

4.3.1. Loss measurements

With proper alignment of the PLA it is possible to make length adjustments of over 15 cm with only a 5% change in optical intensity to the photodiode. This corresponds to a 10% change in the electrical power delivered to the antenna. Path length adjustments of 8 cm or less result in a change in optical intensity of less than 2%. Figure 4.6 shows a plot of the actual intensity change versus the change in path length for each of the four channels. The overall insertion loss is plotted against separation for each PLA in Fig. 4.7.

The greater variability in channel 4 dictates its choice as the reference channel, i.e., the channel for which the path length remains fixed.

4.3.2. Assembly procedure

The collimator consists of a glass sleeve (1) of inside diameter 1.83 mm, the elastomeric subsection of a GTE lab splice (2), and a 0.23 pitch GRIN lens of outside diameter 1.80 mm (3). The fiber is centered in the glass sleeve by the elastomeric splice. The elastomeric splice has been cut in half. The fiber passes through the elastomeric splice and protrudes by about 0.5 mm. It is held firmly in place by the splice. The GRIN lens is then cemented in place once it is positioned in the glass sleeve such that the emerging beam is well collimated.

The actual insertion of the fiber into the elastomeric splice takes place with the splice only partially inserted into the glass sleeve. This reduces the force necessary to insert the fiber and prevents breakage. The splice is then pushed into the sleeve with a pair of tweezers. A sketch of the assembled collimator may be found in Fig. 3.4.

The GRIN lens is focused as follows. A piece of thin double back tape is wrapped over the cutting edge of an Exacto blade, which is then secured to an xyz stage. The cutting edge is gently placed in contact with the cylindrical edge of the GRIN lens. With the tape adhering to the lens the xyz stage is used to move the lens in and out until the beam is collimated. After the final adjustment the lens should protrude from the glass sleeve by several mm. U-V curing epoxy is applied to the lens where it protrudes from the glass sleeve followed by u-v light to cure the epoxy. The epoxy may be applied one drop at a time using a small piece of optical fiber as an applicator.

5. Conclusions

The apparatus described in the above report was designed to investigate the performance of a fiber optic distribution system used for beam steering with a phased array antenna at x-band microwave radiation. Discrete components were used to breadboard the system, parts of which might later be put into integrated format.

A novel design for the fabrication of a fiber optic collimator and path length adjuster was described and test results cited.

Although the apparatus described in this report was designed primarily for the study of beam steering, it has proved to be both versatile and reliable and could find applications in other studies. For example, it could be used to provide multiple electrical (or optical) signals at gigahertz frequencies with infinitely variable time delay from less than one picosecond to several nanoseconds. Time delays could be programmed quickly and easily by mounting the appropriate element of the path length adjuster on a computer driven translation stage. This would also permit automated data acquisition. Moreover, the amplitude of each channel could be adjusted by inserting neutral density filters in the beam within the PLAs. This, too, could be automated using a variable neutral density filter disk.

6. References

1. Glass sleeve samples were provided by Mr. A. Melite, Fiber Optics Division, GTE Products Corporation, Williamsport, PA 17701
2. Lab splices are obtainable from GTE Products Corporation, Williamsport, PA 17701
3. NSG America, Inc., 28 Worlds Fair Drive, Somerset, NJ 08873, part # SLW-1.8-0.23-B2-0.83-0-00

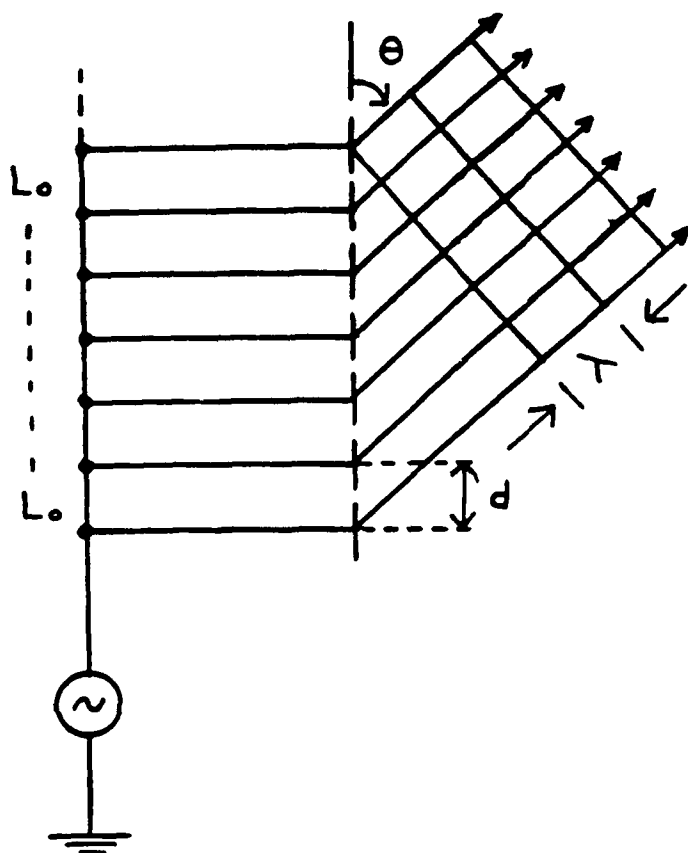


Figure 3.1. Progressive time delay between antenna elements caused by the incremental path difference L_0 results in a progressive phase shift (β). The beam is directed at angle θ .

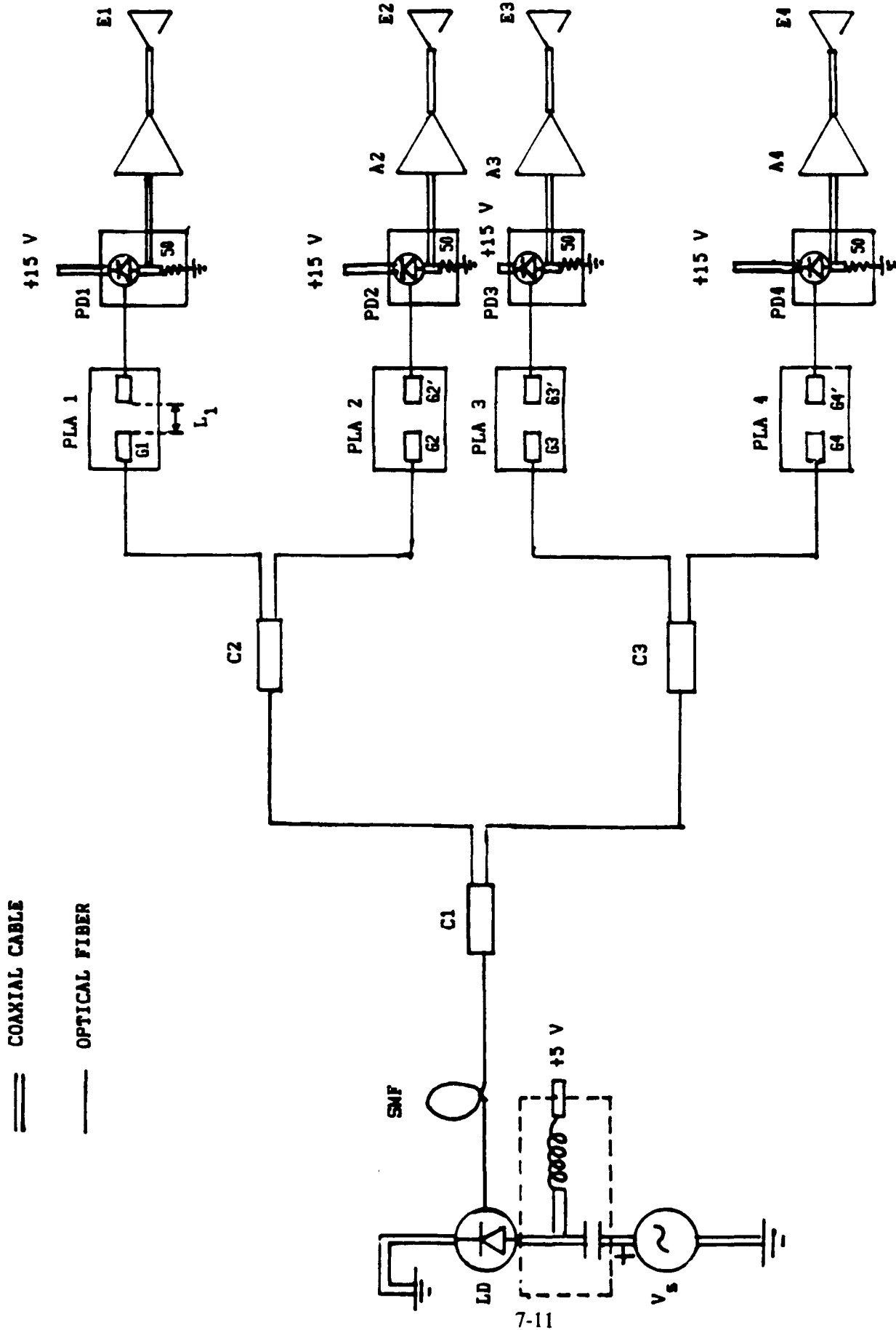


Figure 3.2. Block diagram of a four element phased array antenna.

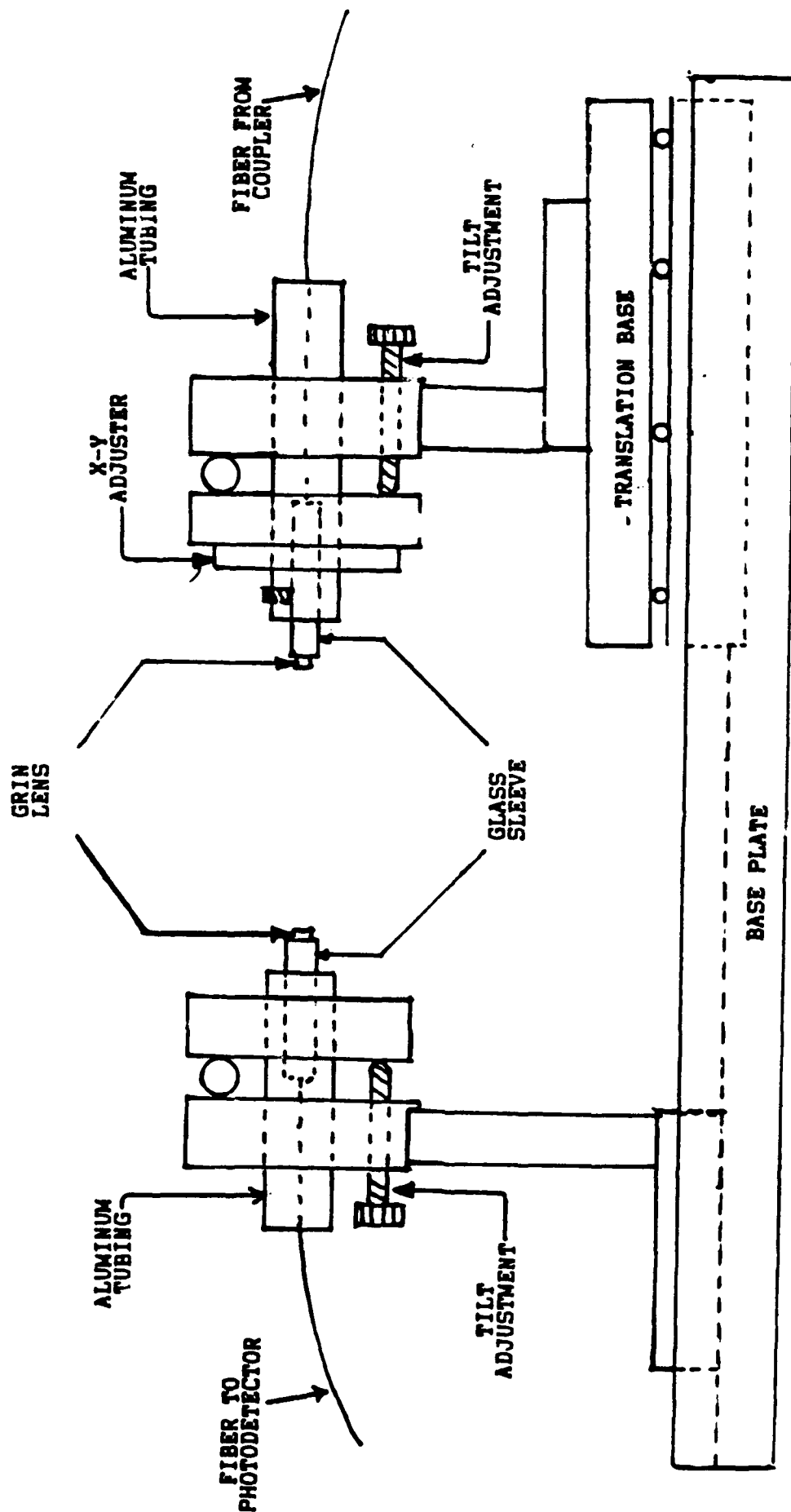


Figure 3.3. Path length adjuster.

Note: The micrometer driven translation base (to the right in the diagram) mounts to the base plate at only one location. The GRIN lens mount (to the left in the diagram) slides in a machined channel and may be positioned at any distance from the translation base up to about 20 cm.

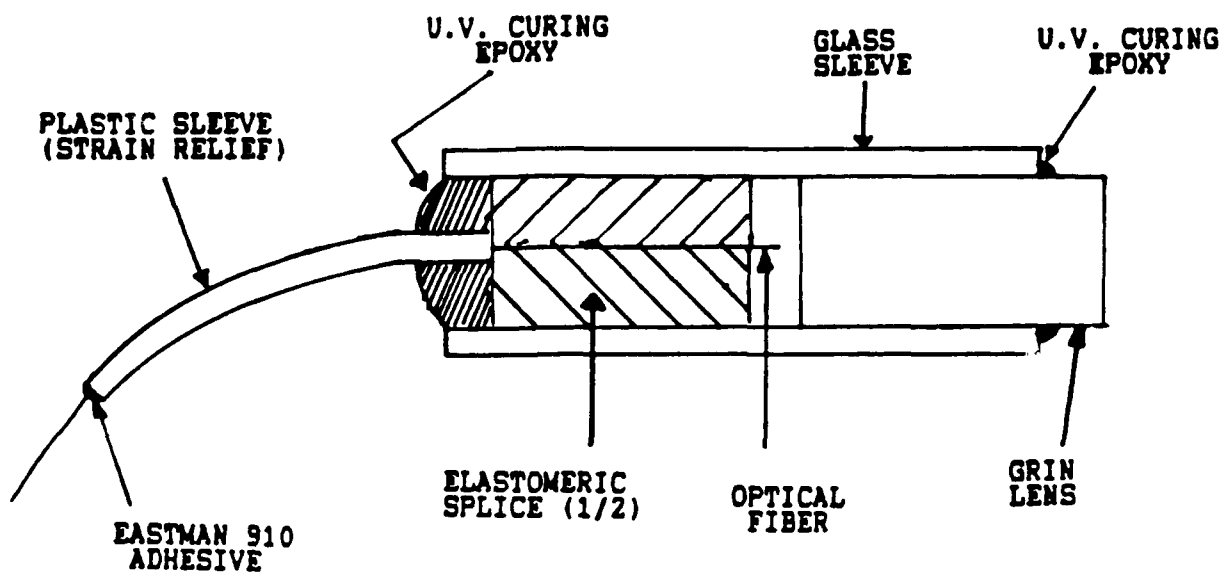


Figure 3.4. GRIN lens collimator.

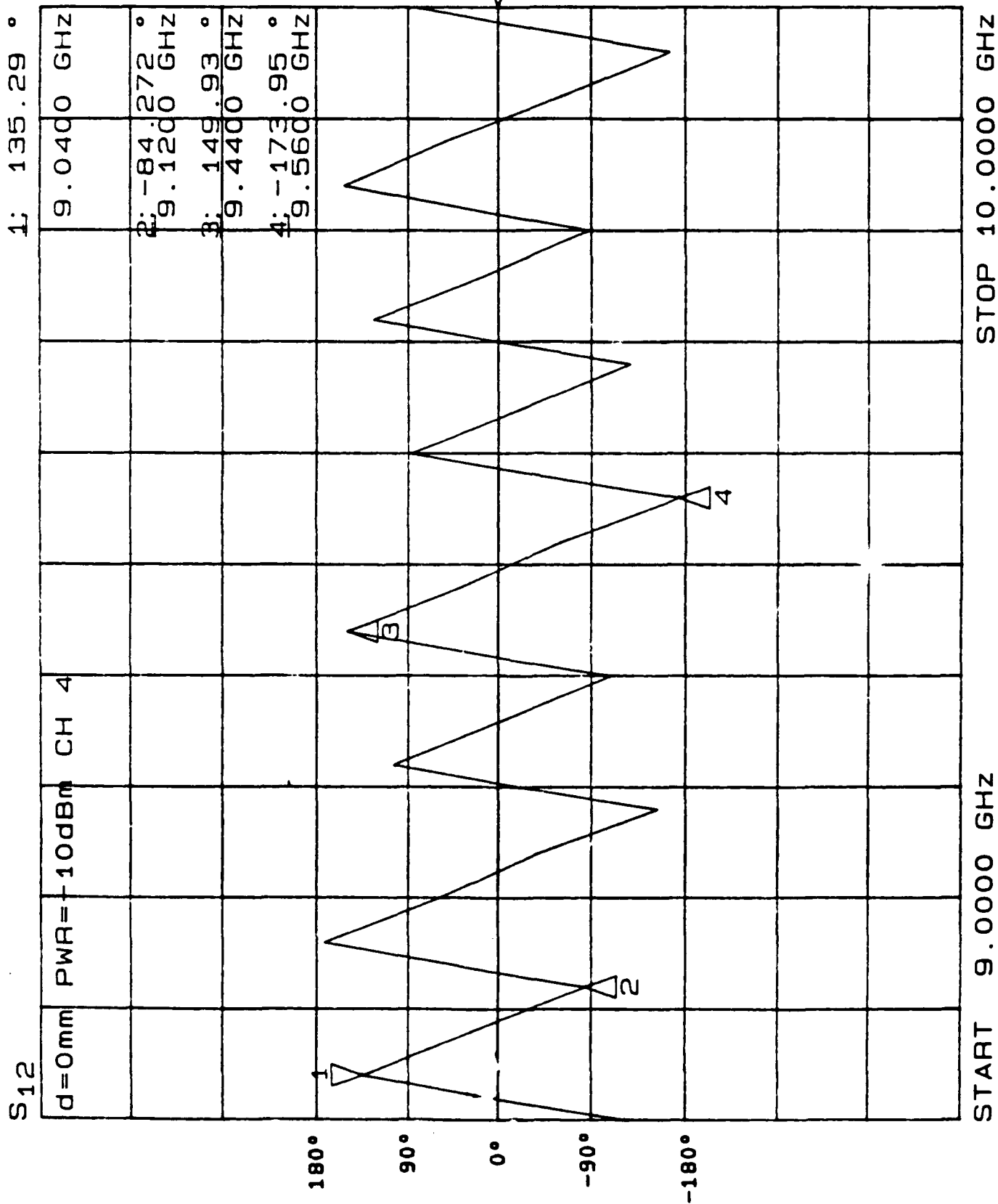


Figure 4.1. Phase vs. frequency for channel 4. Channel 4 is the reference channel.

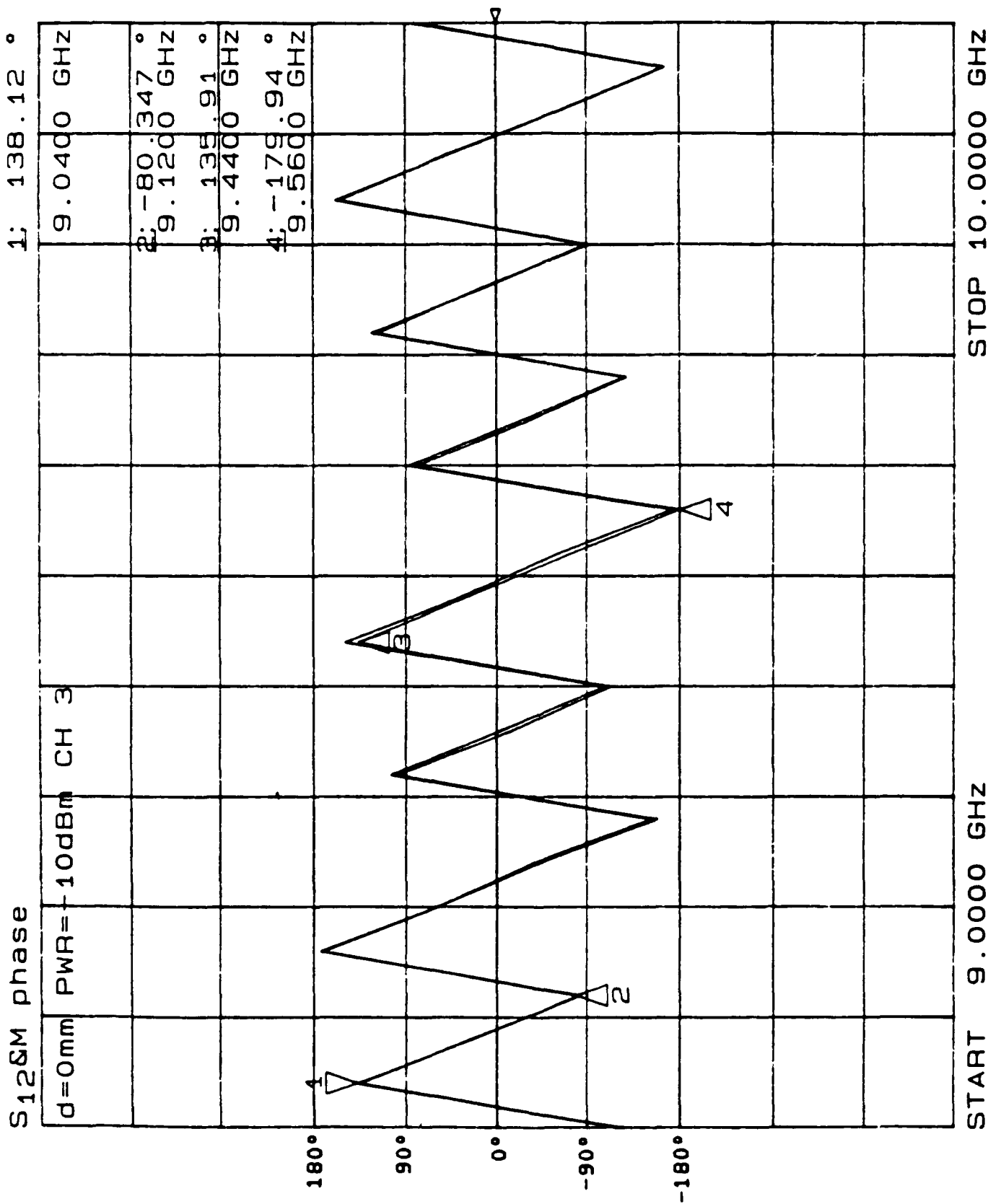


Figure 4.2. Phase vs. frequency for channel 3. Channel 3 is indicated by the markers. Channel 4 is shown as an overlay.

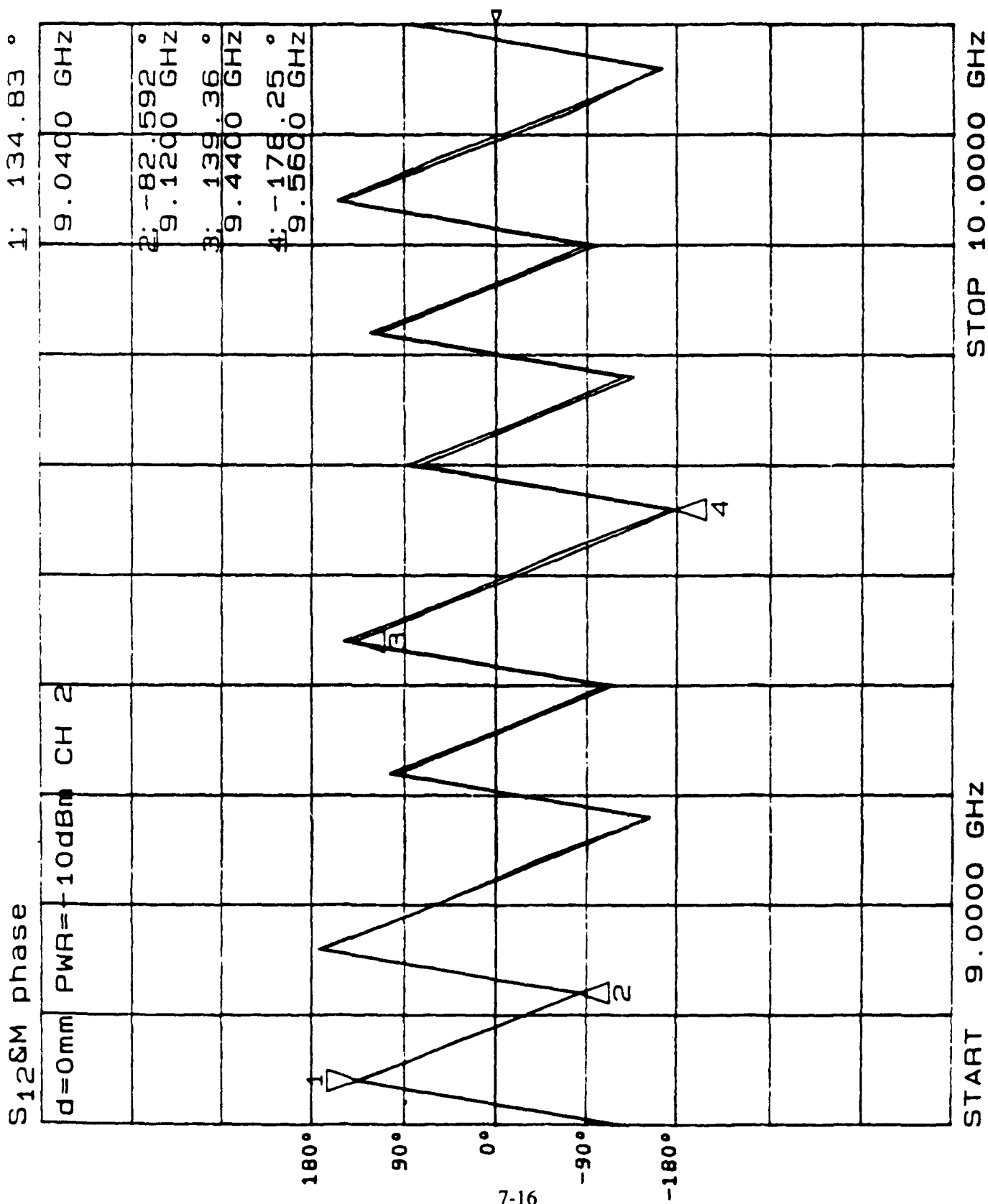


Figure 4.3. Phase vs. frequency for channel 2. Channel 2 is indicated by the markers. Channel 4 is shown as an overlay.

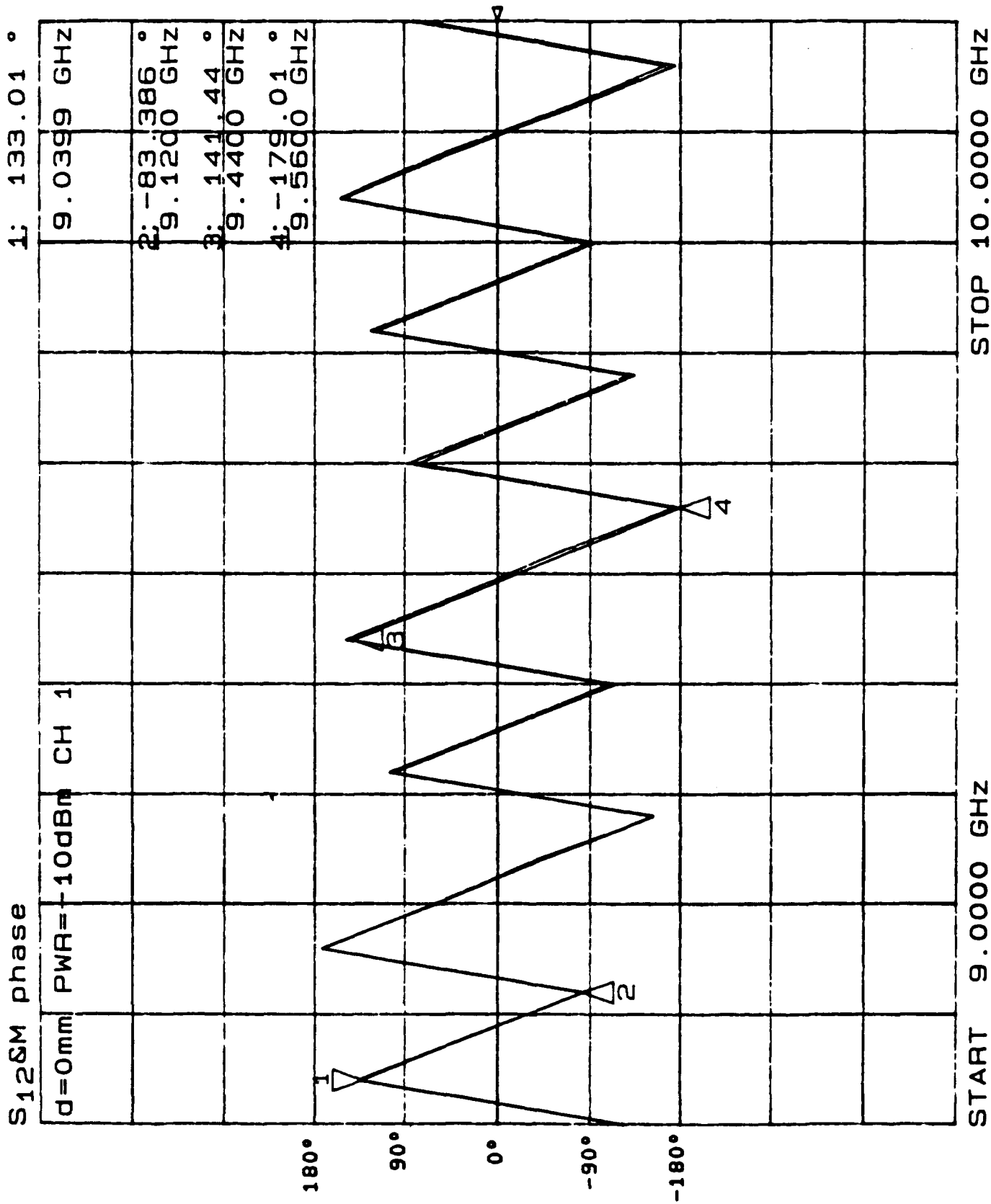


Figure 4.4. Phase vs. frequency for channel 1. Channel 1 is indicated by the markers. Channel 4 is shown as an overlay.

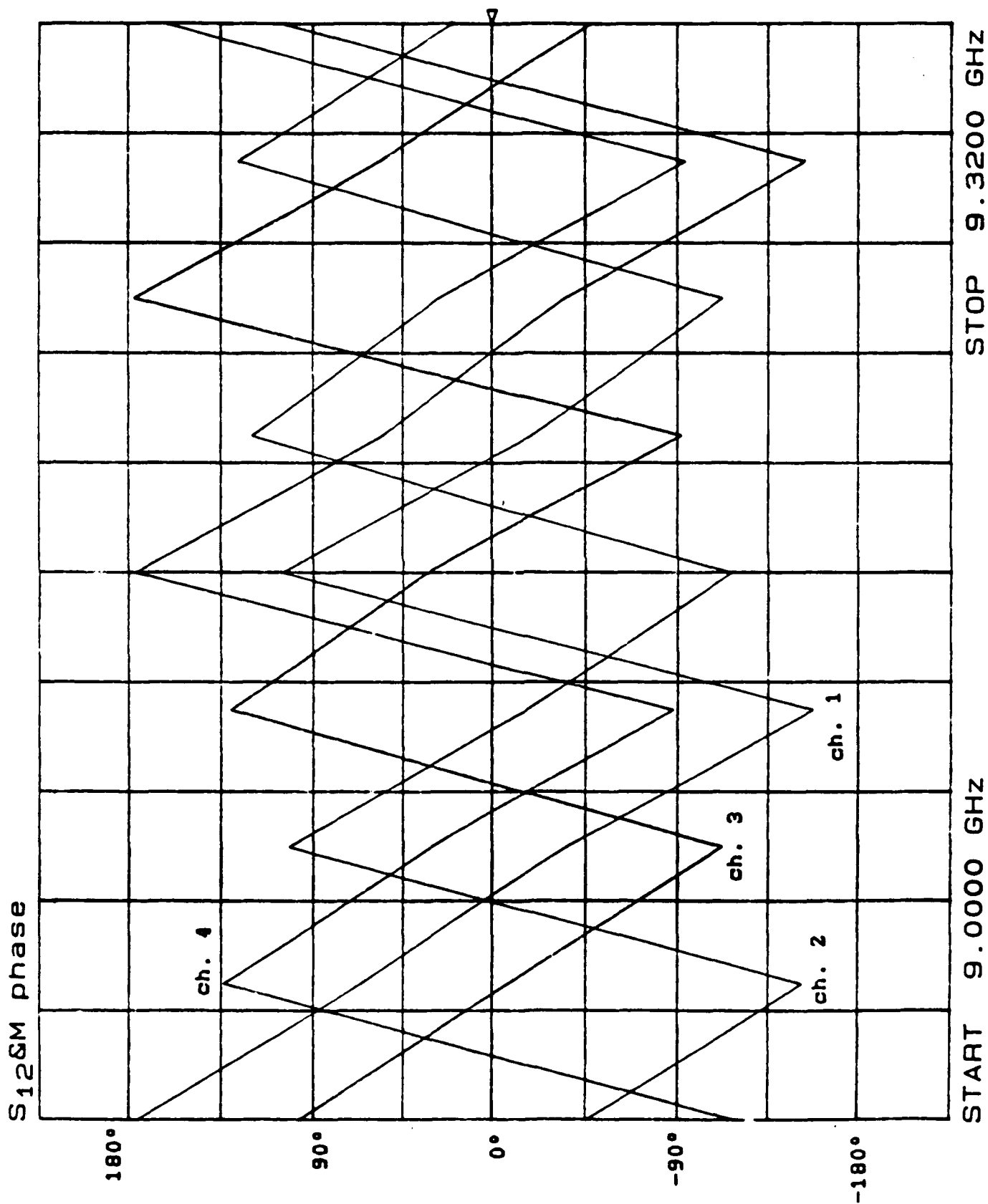


Figure 4.5. Phase shifts vs. frequency for channels 1-4 with the PLAs adjusted for 60° beam steering.

PLA: INTENSITY CHANGE VS. SEPARATION

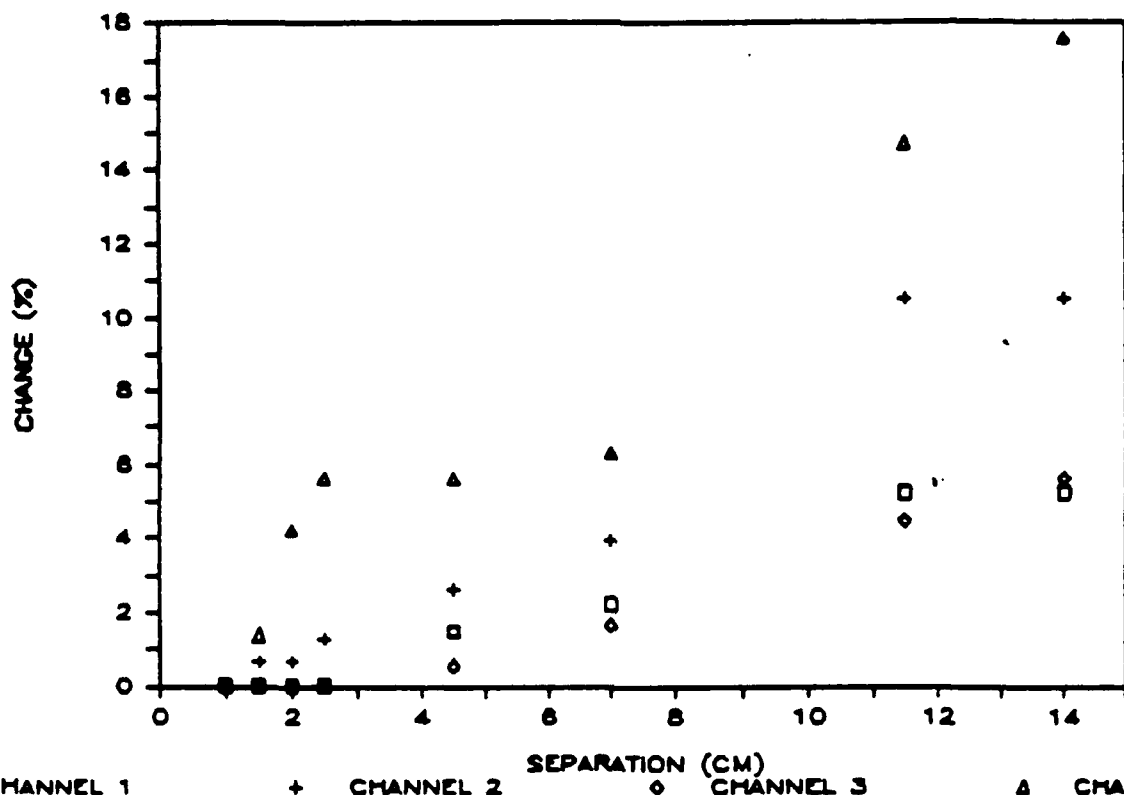


Figure 4.6.

PLA: INSERTION LOSS VS. SEPARATION

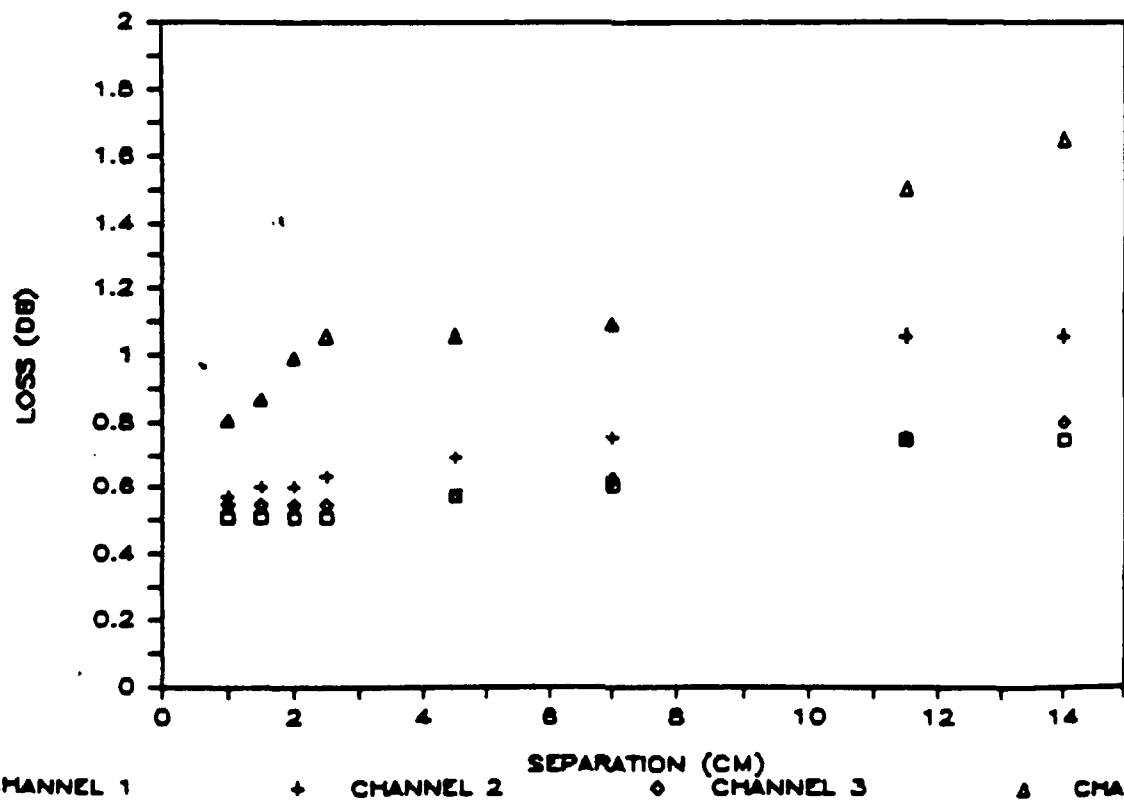


Figure 4.7.

FINAL REPORT

Submitted to
Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432-1984

Continuation Study of a Communications Receiver
for
Spread Spectrum Signals

Prepared by:	Donald R. Ucci, Ph.D.;
	William Jacklin, Jimm Grimm
Academic Rank:	Associate Professor;
	Research Assistants
Department and	Electrical and Computer Engineering Dept.
University:	Illinois Institute of Technology
Research Location:	Rome Laboratories/DCCD
	Griffiss Air Force Base, NY 13441
USAF Researcher	Mr. John Patti and Mr. Steve Tyler
Date	January 31, 1991
Contract No:	F49620-88-C-0053/SB5881-0378
Subcontract No:	S-210-10MG-067

ABSTRACT

This report presents the theoretical development and the simulation of an adaptive nonlinear communications receiver based on an M-Interval Polynomial Approximation (MIPA) of the Probability Density Function (PDF) for the received signal envelope. This system has the potential to perform robustly for a wide range of interference scenarios. Preliminary results for transmission in the presence of Continuous Wave (CW) and Wide Band (WB) interference are presented and discussed. A comparison of this method to another nonlinear processor is performed. The preliminary results indicate that the performance of the various simulations is similar, and that all of the simulations perform substantially better than linear receivers. Recommendations for continued topics of research are also discussed.

ACKNOWLEDGEMENT

We express our thanks and appreciation to the Air Force Systems command, the Air Force Office of Scientific Research, Rome Laboratories, and Universal Energy Systems for making this research project possible. In particular, we thank Mr. John Patti and Mr. Steve Tyler, our research colleagues at Rome Laboratories, Dr. Rodney C. Darrah, director of the Universal Energy Systems Research Initiation Program, and the staff of both these organizations.

1. INTRODUCTION

Many facets are involved in the design of a communications system. One of considerable importance is determining a method to recover the transmitted signal when it is subjected to interference in the transmission path. The first step in the process is choosing a suitable interference model. Using this model, it is possible to design a receiver that will either remove the effects of the interference on the transmitted signal, or operate on the transmitted signal plus interference in such a way as to make determination of the correct transmitted waveform possible. This subject is the focus of the current study.

For many practical applications, it is possible to model interference using a Gaussian Probability Density Function (PDF), given by:

$$f_N(n) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{(n-\mu)^2}{2\sigma^2}} \quad (1-1)$$

where: $N(t)$ is the noise random process,
 μ is the mean of the noise,
 σ^2 is the variance.

The optimum processor for this type of interference is a linear processor which incorporates the use of a matched

filter for signal recovery. The matched filter impulse response is:

$$h(t) = kx(T-t) \quad (1-2)$$

where: $h(t)$ is the matched filter impulse response,
 $x(t)$ is the transmitted signal waveform (real),
 T is the period of the transmitted waveform,
 k is an arbitrary constant.

Alternately, a time correlator realization of the matched filter can be used for signal recovery. The time correlator is derived from examining the output of the matched filter at time $t=T$, which can be written as:

$$y(T) = \int_0^T r(t) \times x(t) dt \quad (1-3)$$

where: $y(T)$ is the output of the matched filter at time
 T ,
 $r(t)$ is the received signal,
 $x(t)$ is the transmitted signal waveform.

Either of these processing techniques can be used by a receiver to recover a transmitted signal in the presence of Gaussian noise. When operating in a two-dimensional environment, these operations must be performed on both the In-phase (I) and Quadrature (Q) channels.

However, in many applications the interference is of a non-Gaussian nature. Examples include Continuous Wave (CW) interference, impulsive interference, and non-Gaussian Wide Band (WB) interference. In these cases it is necessary to employ nonlinear processing techniques to recover the transmitted waveform at the receiver. Examples of such processing have been described in a document by Hazeltine Corporation^[1] and by J. L. Higbie^[2]. Their methods are compared in Appendix A for the case of a two-dimensional signal environment.

The method investigated in this study utilizes the likelihood function of statistical detection theory^[1]. This technique requires determination of the received signal PDF. One procedure which evolved from this approach utilizes the Locally Optimal (LO) maximum likelihood function^[1]. For a two-dimensional receiver, the LO likelihood function is derived in Appendix B and has the form:

Choose \vec{r}_j and \vec{s}_j , $j=1 \rightarrow N$, to maximize

$$\bar{L}_j = \sum_{k=1}^K \{r_{kj} \times \cos \theta \times g(r) + s_{kj} \times \sin \theta \times g(r)\} \quad (1-4)$$

with,

$$g(r) = - \frac{\frac{d}{dr} f_e(r)}{f_e(r)} + \frac{1}{r} \quad (1-5)$$

where: \bar{L}_j is the LO maximum likelihood function for the j^{th} transmitted signal pair,

$f_e(r)$ is the received envelope PDF,

\bar{r}_j, \bar{s}_j are the transmitted signal vectors for the I and Q channel respectively,

r_{kj}, s_{kj} is the k^{th} of K samples used to represent the j^{th} transmitted signal pair,

N is the number of total possible transmitted signal pairs in the signal space,

K is the length of a signal vector, i.e. the number of samples per symbol,

r is the received magnitude,

θ is the received phase angle,

$g(r)$ is the optimum nonlinearity.

In practice, however, it is difficult to construct a continuous PDF or to take a continuous derivative of a discrete function. Therefore, it is necessary to use a method of approximation to develop the optimum nonlinearity, $g(r)$. Two methods described by Hazeltine Corporation^[1] are examined in this study: the histogram approximation and the M-Interval Polynomial Approximation (MIPA). In statistics, the histogram method is the classical approach for approximating a PDF. The MIPA algorithm is an extension of the histogram approach and utilizes a set of polynomial functions to approximate the PDF. In the case at hand, this set of functions as a whole

describes completely the entire PDF of the received signal envelope. An added benefit to this approach is that the derivative of a polynomial is easily calculated. Therefore, this algorithm allows for approximation of the optimum nonlinearity in real time. Notice that a zero-order MIPA reduces to a histogram. Thus, these two methods are interrelated.

In addition to examining the MIPA implementation of the LO maximum likelihood function, a nonlinear processing technique developed by J. L. Higbie^[2] was examined, and a preliminary comparison to the MIPA algorithm was made. The Higbie approach was implemented by Rome Laboratories (RL) (formerly Rome Air Development Center (RADC)) in the current Adaptive Nonlinear Coherent Processor (ANCP) simulation. A theoretical comparison between the LO and Higbie optimum nonlinearities is presented in Appendix A.

The development of the MIPA algorithm and the implementation of the LO maximum likelihood function have comprised the major efforts of the current study. The following sections present the results and conclusions of the simulation of these techniques, as well as comparisons to the Higbie approach. Also, because of discrepancies in the Hazeltine report, the derivation of the LO maximum likelihood function for two-dimensional signals is presented in Appendix B.

2. OBJECTIVES

The ultimate goal of this research is to determine an effective means for mitigating the effects of non-Gaussian interference in a communications system. To achieve this task many intermediate goals must be accomplished. As stated in the proposal for the current project, these include:

- 1) Examine the M-Interval Polynomial Approximation (MIPA) approach to modeling a Probability Density Function (PDF). Extend this theory to a two dimensional, i.e. in-phase and quadrature channel, receiver.
- 2) Integrate the MIPA implementation of the Locally Optimal (LO) maximum likelihood function into the existing Adaptive Nonlinear Coherent Processor (ANCP) simulation.
- 3) Compare the performance of the MIPA to the histogram approximation of the PDF.
- 4) Determine the optimal order of polynomial and number of intervals for the MIPA.
- 5) Incorporate Bit Error Rate (BER) as a performance metric.

- 6) Test the system under various forms of modulation such as Offset Quaternary Phase Shift Keying (OQPSK) and Minimum Shift Keying (MSK).
- 7) Optimize the simulation for efficient and fast performance.

The current status of each intermediate task is:

- 1) Development of the MIPA algorithm for approximating the received envelope PDF was accomplished.
- 2) Integration of the MIPA implementation of the LO receiver into the ANCP simulation was accomplished.
- 3) Preliminary comparison between the MIPA and the histogram approximation of the received envelope PDF was performed.
- 4) Preliminary comparisons concerning the optimum order and number of intervals for the MIPA algorithm was performed. More investigation is necessary before conclusions can be formulated.
- 5) Incorporation of BER analysis will require extensive modification of the current simulation. This task will be addressed in future research.

- 6) Through consultation with Rome Laboratories (RL) it was determined that studying MSK in detail would be more beneficial than examining many different modulation schemes.
- 7) Preliminary methods for optimization were used in the MIPA portion of the current simulation. More efforts are required to optimize the entire ANCP simulation.

3. THEORETICAL DEVELOPMENT OF THE M-INTERVAL POLYNOMIAL APPROXIMATION (MIPA) NONLINEAR PROCESSOR

3.1 DEVELOPMENT OF THE LOCALLY OPTIMAL MAXIMUM LIKELIHOOD FUNCTION FOR TWO-DIMENSIONAL SIGNALS

In modern communications systems, it is necessary to develop methods to improve signal reception when interference signals are present in the transmission path. The use of nonlinear signal processing based on the likelihood function is one such method. In particular, the focus of the current study is on developing the likelihood function for complex, or two-dimensional, signals and then using an M-Interval Polynomial Approximation (MIPA) Algorithm for implementation. A full development of the likelihood function is presented in Appendix B, but the key steps are shown here.

The transmitted, or source, signal is represented by a random process with an in-phase component $R(t)$ and a quadrature component $S(t)$. For the received signal, the in-phase component is $I(t)$ and the quadrature component is $Q(t)$. Finally, the interference, or noise, is characterized by an in-phase signal $X(t)$, and a quadrature signal $Y(t)$. The received signal is sampled once per symbol period. The resulting relationship is:

$$I=R+X \quad Q=S+Y \quad (3-1)$$

The received signal joint PDF is given by:

$$f_{IQ}(i, q) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{RS}(r, s) f_{XY}(i-r, q-s) dr ds \quad (3-2)$$

where: $f_{RS}(r, s)$ is the joint PDF of the transmitted signal,

$f_{XY}(x, y)$ is the joint PDF of the interference.

If the transmitted signal is one of N equiprobable transmitted signal pairs, then Eq. (3-2) reduces to:

$$f_{IQ}(i, q) = \frac{1}{N} \sum_{j=1}^N f_{XY}(i-r_j, q-s_j) \quad (3-3)$$

where: r_j and s_j are one of N possible transmitted signal pairs.

For the likelihood function, it is desired to maximize $P(R, S|I, Q)$, the probability that $R=r_j$ and $S=s_j$ given that $I=i$ and $Q=q$, where i and q are known. This is equivalent to choosing r_j and s_j to maximize $f_{XY}(i-r_j, q-s_j)$.

If each signal pair is sampled to form vectors of length K , i.e. $\vec{I} = \vec{R} + \vec{X}$ and $\vec{Q} = \vec{S} + \vec{Y}$, Eq. (3-3) must be modified. When independent and identically distributed samples are assumed, the corresponding likelihood function becomes:

Choose \vec{r}_j and \vec{s}_j , $j=1 \rightarrow N$, to maximize:

$$L_j' = \prod_{k=1}^K f_{XY}(i_k - r_{jk}, q_k - s_{jk}) \quad (3-4)$$

where: K is the length of a signal vector,
 i_k, q_k is the k^{th} sample of the received signal
for the I and Q channels respectively.

To simplify calculations, the natural logarithm of Eq. (3-4) is used. This results in the Globally Optimal (GO) maximum likelihood function:

Choose \vec{r}_j and \vec{s}_j , $j=1 \rightarrow N$, to maximize

$$L_j = \sum_{k=1}^K \ln[f_{XY}(i_k - r_{jk}, q_k - s_{jk})] \quad (3-5)$$

For a large Jammer to Signal (J/S)¹ ratio, a first-order Taylor Series expansion around the received signal point can be used to approximate the interference PDF. The expansion is valid because the deviation of the interference from the received signal point is minimal for a large J/S. This simplification results in the Locally Optimal (LO) maximum likelihood function:

¹A typical jammer is usually at least 20dB greater than the transmitted signal. Therefore, for a signal environment where jammers are present, a large J/S assumption is valid.

Choose \bar{r}_j and \bar{s}_j , $j=1 \rightarrow N$, to maximize

$$\bar{L}_j = - \sum_{k=1}^K \left\{ r_{kj} \times \frac{\frac{\partial}{\partial i} f_{IQ}(i_k, q_k)}{f_{IQ}(i_k, q_k)} + s_{kj} \times \frac{\frac{\partial}{\partial q} f_{IQ}(i_k, q_k)}{f_{IQ}(i_k, q_k)} \right\} \quad (3-6)$$

where: $f_{IQ}(\cdot)$ is the received signal PDF.

However, because of the limited ability of the simulator to approximate continuous functions, determination of the received signal joint PDF can be a difficult process, and determination of the partial derivatives even more so. To reduce the complexity of the likelihood function, complex bivariate radial symmetry of the received signal PDF is assumed. Therefore, the received signal PDF is given by:

$$f_{IQ}(i, q) = \begin{cases} \frac{f_e(r)}{2\pi r}, & 0 < \theta < 2\pi \\ 0, & \text{elsewhere} \end{cases} \quad (3-7)$$

where: $f_e(r)$ is the received envelope PDF,
 $r = \sqrt{i^2 + q^2}$ is the received magnitude,
 $\theta = \arctan \frac{q}{i}$ is the received phase angle.

This assumption is valid because interference sources of interest will not have predictable phase angles. Even a constant-frequency waveform will have a vector that rotates at a constant rate and is therefore equally likely at any angle.⁽²⁾ With this assumption the likelihood function

reduces to Eq. (1-4), repeated below:

Choose \vec{r}_j and \vec{s}_j , $j=1 \rightarrow N$, to maximize

$$\bar{L}_j = \sum_{k=1}^K \left\{ r_{kj} \times \cos \theta \times \left[-\frac{\frac{d}{dr} f_e(r)}{f_e(r)} + \frac{1}{r} \right] + s_{kj} \times \sin \theta \times \left[-\frac{\frac{d}{dr} f_e(r)}{f_e(r)} + \frac{1}{r} \right] \right\} . \quad (3-8)$$

This is the likelihood function that is implemented in the ANCP simulation. Of importance is the optimum nonlinearity of Eq. (1-5), also repeated below for convenience:

$$g(r) = -\frac{\frac{d}{dr} f_e(r)}{f_e(r)} + \frac{1}{r} \quad (3-9)$$

The MIPA algorithm is one method by which the optimum nonlinearity can be computed.

3.2 DEVELOPMENT OF THE MIPA ALGORITHM^[1]

The nonlinear function expressed in Eq. (3-9) requires knowledge of the probability density function (PDF) of the received signal. Since it is difficult to determine the exact value of the PDF, an approximation must be made. The two approximations investigated are the histogram approximation and the M-Interval Polynomial Approximation (MIPA)^[1].

The acronym MIPA requires further explanation. The term M-Interval is used because the received signal PDF is divided into M intervals. The Polynomial Approximation term is used because the PDF is subsequently approximated by a separate polynomial over each interval. Thus, a MIPA is simply a concatenation of polynomial curves used to approximate the actual PDF of the received signal. MIPAs of order 0, 2, and 4 are investigated in this report, where the order refers to the order of the polynomials used. It turns out that the zero order MIPA reduces to a histogram since zero order polynomials are constants which produce horizontal lines. The boundaries, or breakpoints of each interval are determined by:

$$h_m = \frac{V_{\max} - V_{\min}}{M}, m=1, 2, \dots, M \quad (3-10)$$

where: h_m is the value of the m^{th} breakpoint,

V_{\max} is the maximum value of the PDF,

V_{\min} is the minimum value of the PDF,

M is the number of intervals.

The symbols α and β denote the lower and upper breakpoints, respectively, of the interval of interest.

Since a separate polynomial is used to approximate each interval, the PDF of the received signal conditioned on the interval of interest is required. This is written as:

$$f_{ec}(r) = f_e(r|\alpha \leq r < \beta) = \frac{f_e(r)}{P(\alpha \leq r < \beta)} \quad (3-11)$$

where: $f_e(r)$ is the received envelope PDF,
 $P(\alpha \leq r < \beta)$ is the probability that r is in the
interval $[\alpha, \beta]$.

The general form of the approximating polynomial is:

$$\hat{f}_{ec}(r) = \sum_{k=0}^P b_k r^k \quad (3-12)$$

where: $\hat{f}_{ec}(r)$ is the MIPA of $f_{ec}(r)$,
 P is the order of the polynomial,
 b_k are the polynomial coefficients.

Values for b_k must be determined such that the polynomial best approximates the PDF according to some error criteria. A common and tractable method is the least squares error minimization^[3]. Applying this yields:

$$\varepsilon = \int_a^B (f_{ec}(r) - \hat{f}_{ec}(r))^2 dr \quad (3-13)$$

where: ε is the error to be minimized.

The general form of the approximating polynomial from Eq. (3-12) can be substituted for $\hat{f}_{ec}(r)$ into Eq. (3-13).

$$\varepsilon(B) = \int_a^B (f_{ec}(r) - \sum_{k=0}^P b_k r^k)^2 dr \quad (3-14)$$

where: $\varepsilon(B)$ is the error as a function of B ,

B is the polynomial coefficient vector $[b_0 \ b_1 \dots b_p]$.

The error can be minimized for each coefficient by taking the derivative of $\varepsilon(b_i)$ with respect to each b_i , setting it equal to zero, and solving for b_i .

$$\frac{d\varepsilon(b_i)}{db_i} = \int_a^B -2r^i (f_{ec}(r) - \sum_{k=0}^P b_k r^k) dr = 0, \quad i=0, 1, \dots, P \quad (3-15)$$

Simplifying Eq. (3-15) yields

$$\int_{\alpha}^{\beta} f_{ec}(r) r^i dr = \int_{\alpha}^{\beta} r^i \sum_{k=0}^p b_k r^k dr \quad (3-16)$$

The term on the left side of Eq. (3-16) is just the i^{th} moment of $f_{ec}(r)^{[3]}$, i.e. m_{xi} . Solving the integral on the right side yields:

$$m_{xi} = \frac{\sum_{k=0}^p b_k (\beta^{i+k+1} - \alpha^{i+k+1})}{i+k+1}, \quad i=0, 1, \dots, P \quad (3-17)$$

Equation (3-17) can be written in matrix form by letting i be the row index and k be the column index. Thus, MIPA coefficients of general order are given by:

$$\begin{bmatrix} m_{x0} \\ m_{x1} \\ \vdots \\ m_{xP} \end{bmatrix} = \begin{bmatrix} \frac{\beta - \alpha}{1} & \frac{\beta^2 - \alpha^2}{2} & \dots & \frac{\beta^{P+1} - \alpha^{P+1}}{P+1} \\ \frac{\beta^2 - \alpha^2}{2} & \frac{\beta^3 - \alpha^3}{3} & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\beta^{P+1} - \alpha^{P+1}}{P+1} & \dots & \dots & \frac{\beta^{2P+1} - \alpha^{2P+1}}{2P+1} \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_P \end{bmatrix} \quad (3-18)$$

For notational purposes, Eq. (3-18) can be written as:

$$\vec{m}_x = H_x \times \vec{b} \quad (3-19)$$

$$\vec{b} = H_x^{-1} \times \vec{m}_x \quad (3-20)$$

where: \vec{m}_x is the $(P+1) \times 1$ vector of moments,

H_x is the $(P+1) \times (P+1)$ matrix of intervals.

The vector \vec{b} can be computed directly from Eq. (3-20) since \vec{m}_x can be computed from the received signal, α and β can be computed from M , and P is a parameter of the simulation. However, this method requires the inversion of H_x , which is a time consuming task. The performance of the simulator can be improved significantly by avoiding this computation.

The matrix relationships shown in Eq. (3-19) and Eq. (3-20) can be normalized by mapping the interval $[\alpha, \beta]$ onto the interval $[0, 1]$ using the linear mapping:

$$y = \frac{x - \alpha}{\beta - \alpha} \quad (3-21)$$

Under this mapping,

$$H_y = \begin{bmatrix} \frac{1}{1} & \frac{1}{2} & \dots & \frac{1}{P+1} \\ \frac{1}{2} & \frac{1}{3} & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{1}{P+1} & \dots & \dots & \frac{1}{2P+1} \end{bmatrix} \quad (3-22)$$

Note from Eq. (3-22) that H_y is a constant, and therefore so is H_y^{-1} . Thus, the inverse of H_y need only be computed once when the receiver is initially activated. Without the linear mapping, H_x^{-1} would need to be computed for every interval of the PDF. The moment vector \vec{m}_y of the random variable Y may be determined using the following equations^[3]:

$$m_{yi} = \int_{-\infty}^{\infty} y^i f_{yc}(y) dy \quad (3-23)$$

or,

$$m_{yi} = \int_{-\infty}^{\infty} \left(\frac{r-\alpha}{\beta-\alpha} \right)^i f_{ec}(r) dr \quad (3-24)$$

where: $f_{ec}(r)$ is $f_e(r|\alpha \leq r < \beta)$, the conditional signal PDF with r as index,

$f_{yc}(y)$ is $f_y(y|\alpha \leq y < \beta)$, the conditional signal PDF with y as index,

m_{yi} is the i^{th} moment of $f_{yc}(y)$.

Equation (3-24) can be written in matrix form by letting i be the row index:

$$\vec{m}_y = \begin{bmatrix} 1 \\ \frac{m_{x1} - \alpha}{(\beta - \alpha)} \\ \frac{m_{x2} - 2\alpha m_{x1} + \alpha^2}{(\beta - \alpha)^2} \\ \frac{m_{x3} - 3\alpha m_{x2} + 3\alpha^2 m_{x1} - \alpha^3}{(\beta - \alpha)^3} \\ \vdots \end{bmatrix} \quad (3-25)$$

Equations (3-22) and (3-25) can be used to find

$$\vec{d} = H_y^{-1} \times \vec{m}_y \quad (3-26)$$

The normalized coefficient vector \vec{d} is an intermediate vector used for computational efficiency. A transition matrix Q is necessary to convert \vec{d} to B , i.e.

$$B = Q \times \vec{d} \quad (3-27)$$

The required Q is:

$$Q = \begin{bmatrix} \frac{1}{(\beta - \alpha)} & \frac{-\alpha}{(\beta - \alpha)^2} & \frac{\alpha^2}{(\beta - \alpha)^3} & \dots & \frac{(-1)^P \alpha^P}{(\beta - \alpha)^{(P+1)}} \\ 0 & \frac{1}{(\beta - \alpha)^2} & \frac{-2\alpha}{(\beta - \alpha)^3} & & \vdots \\ 0 & 0 & \frac{1}{(\beta - \alpha)^3} & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{(\beta - \alpha)^{(P+1)}} \end{bmatrix} \quad (3-28)$$

Although Q must be calculated for every interval, it need not be inverted, thus substantial computational savings result.

The final values necessary for use in the simulation are scaled values of \tilde{B} . The following scaling removes the conditional nature of the PDF:

$$\tilde{a} = P(\alpha \leq r < \beta) \tilde{B} \quad (3-29)$$

where: $\tilde{a} = [a_0 \ a_1 \ \dots \ a_p]$ are the coefficients for $\hat{f}_e(r)$. Thus, the simulation performs the following computation for each interval:

$$\hat{f}_e(r) = \sum_{k=0}^p a_k r^k \quad (3-30)$$

where: $\tilde{a} = Q \times H_Y^{-1} \times \tilde{m}_Y \times P(\alpha \leq r < \beta)$.

The calculation of the optimum nonlinearity and implementation of the LO maximum likelihood function requires the derivative of $f_e(r)$. This derivative is formed by taking the derivative of the approximating polynomial of $f_e(r)$ for each interval. Since each interval of the received envelope PDF is represented by functions of the form:

$$\hat{f}_e(r) = a_0 + a_1 r + \dots + a_p r^p \quad (3-31)$$

the derivative is given by functions of the form:

$$\frac{d}{dr} \hat{f}_e(r) = g_1 + g_2 r + \dots + g_p r^{p-1} \quad (3-32)$$

where: $g_i = i \times a_i$, $i=1, \dots, p-1$

This indexing results in efficient looping algorithms in the simulation.

Therefore, the optimum nonlinearity given by Eq. (3-9) is approximated using the expression:

$$\hat{g}(r) = - \frac{g_1 + g_2 r + \dots + g_p r^{p-1}}{a_0 + a_1 r + \dots + a_p r^p} + \frac{1}{r} \quad (3-33)$$

4. SIMULATION OF MIPA NONLINEAR PROCESSOR

This section discusses the implementation of the MIPA nonlinear processor in the current Adaptive Nonlinear Coherent Processor (ANCP) simulation. The purpose of this discussion is to illustrate the key considerations and data flow necessary for this algorithm. A pictorial view of the MIPA nonlinear processor is shown in Fig. (4-1).

MIPA NONLINEAR PROCESSOR BLOCK DIAGRAM

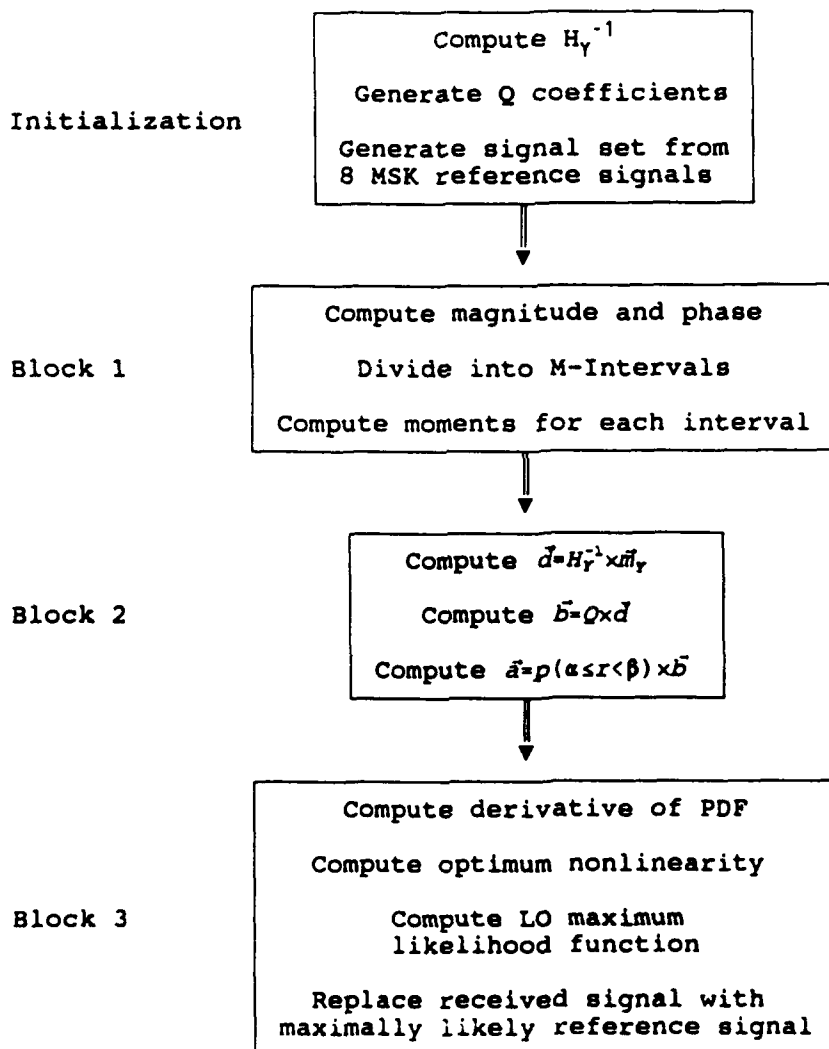


Figure (4-1)

The current simulation supports zeroth (histogram), second and fourth order MIPA approaches. The user inputs are: 1) enable MIPA processing, 2) provide the order of MIPA approximation, and 3) provide the number of intervals used for approximation. During simulator initialization it is necessary to create the constructs that are stored in the nonlinear processor and used in each successive decision cycle. These include: the inverse coefficient calculation matrices, H_y^{-1} , used for computation of the normalized MIPA interval coefficients; the coefficients of elements in the transition matrix, Q , used for conversion between normalized and actual MIPA interval coefficients; and the eight MSK reference transmitted waveforms^[5].

Of interest is the transition matrix, Q , given by Eq. (3-28). As can be seen, each column contains constants multiplied by $\frac{\alpha^{j-i}}{(\beta-\alpha)^{j+1}}$, where j is the column index and i is the row index. With this symmetry, it is possible to construct a matrix consisting of the constant coefficients of the Q matrix. During simulation execution, an efficient algorithm is used which multiplies each successive column of this matrix by $\frac{1}{(\beta-\alpha)}$ and α^{j-i} , accordingly, to compute the actual transition matrix. This eliminates the need for exponentiation, a computationally expensive operation.

Once this information is stored in memory, it is then possible to begin execution of the nonlinear receiver by invoking the first block of the MIPA processor. Since the optimum nonlinearity utilizes the envelope PDF of the received signal, $f_e(r)$, the first step is to calculate the magnitude and phase of the received signal. With this information it is then possible to create a number of equal length partitions for local approximation of $f_e(r)$. The number of partitions is a user defined input. Next, the probability of each interval is determined and the moments for each interval are calculated. The moments are approximated using the equation:

$$m_{Xi} = \frac{1}{N_m} \sum_{\xi=1}^{N_s} r_{\xi}^i \quad (4-1)$$

where: m_{Xi} is the i^{th} moment, $i=1, \dots, P$, of the m^{th} interval, $m=1, \dots, M$,

N_m is the number of received samples that lie between α and β of the m^{th} interval.

If a second order MIPA is chosen, the first and second moments are calculated. For a fourth order MIPA, the first through fourth moments are calculated. Next, the normalized interval moments are calculated using Eq. (3-25). The normalized interval moments are used to determine the MIPA interval coefficients, represented by the vector \vec{a} . These coefficients are the polynomial coefficients used to locally approximate $f_e(r)$.

The second block of the MIPA processor calculates the received envelope PDF approximation, $\hat{f}_e(r)$. This is accomplished in three steps. First, the normalized MIPA interval coefficients are calculated. This is done using Eq. (3-26). Next, Eq. (3-27) is used to calculate the approximating coefficients for the received envelope PDF conditioned on the current interval, $\hat{f}_e(r|(\alpha \leq r < \beta))$. The third step is the calculation of the actual MIPA coefficients used for approximation of $f_e(r)$. This is accomplished by scaling each of the coefficient vectors of $\hat{f}_e(r|(\alpha \leq r < \beta))$ by the probability of the corresponding interval, as in Eq. (3-29).

The calculation of the optimum nonlinearity and implementation of the LO maximum likelihood function are accomplished in the third block of the MIPA processor. The first stage of this block is the calculation of the derivative of $\hat{f}_e(r)$ using Eq. (3-32). The second stage of this block performs simultaneous calculation of the optimum nonlinearity for each received magnitude, and calculation of the likelihood function for each possible transmitted waveform. First, the optimum nonlinearity is calculated for the current received magnitude sample using Eq. (3-33). Then, the contribution of this sample to the likelihood function of each possible transmitted waveform is calculated using:

$$\bar{I}_{kj} = r_{kj} \times \cos \theta \times [g(r)] + s_{kj} \times \sin \theta \times [g(r)] \quad (4-2)$$

$$\text{where: } \bar{L}_j = \sum_{k=1}^K \bar{I}_{kj}.$$

This calculation is performed for a set of K received magnitude samples. Once completed, the largest likelihood function value is chosen, and the corresponding reference waveform is used to replace the received I and Q channel signals. The reference waveforms used are those that were stored in the processor during initialization. This process continues until all of the received signal data is exhausted.

The histogram implementation of the LO maximum likelihood function is currently constructed in parallel with the p^{th} order MIPA implementation. They are quite similar, except in the formulation of the optimum nonlinearity. For the histogram approach, it is difficult to take a continuous derivative because of its discrete nature. By noticing that:

$$-\frac{\frac{d}{dr}f_e(r)}{f_e(r)} + \frac{1}{r} = -\frac{d}{dr}[\ln(f_e(r))] + \frac{1}{r} \quad (4-3)$$

it is possible to estimate the histogram nonlinearity using the relation:

$$g(r) = \frac{-\ln(f_e(r_{m+1})) + \ln(f_e(r_{m-1}))}{r_{m+1} - r_{m-1}} + \frac{1}{r_m} \quad (4-4)$$

where: r_m , $m=1, \dots, M$, is the value of the interval nearest in magnitude to the current received signal sample.

This approximation is used to calculate the likelihood function for each reference waveform. As in the MIPA implementation, the transmitted signal waveform with the largest likelihood function value is used to replace the received I and Q signals.

In either case, the resulting received signal waveform is passed to the rest of the ANCP simulation for demodulation and decoding. The next section illustrates the results of the MIPA and histogram implementations.

5. RESULTS

The Adaptive Nonlinear Coherent Processor (ANCP) simulation uses the MSK encoded message shown in Fig. (5-1) for each trial. The entire frequency spectrum of this message (generated from a 1024 point Fast Fourier Transform) is shown in Fig. (5-2). This spectrum shows both magnitude and phase, and was verified using PC-MATLAB¹. To compare Fig. (5-2) to known MSK spectra^{[4][5]}, a scatter plot in decibels (dB) of the lower frequency components for this spectrum shifted to baseband is shown in Fig. (5-3). This spectrum has been normalized by dividing all frequency components by the highest magnitude.

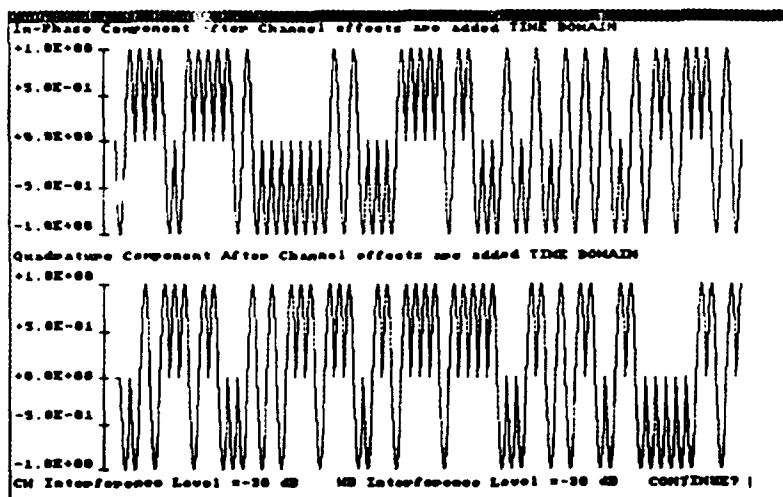


Figure 5-1

¹PC-MATLAB is a registered trademark of The MATH WORKS Inc.

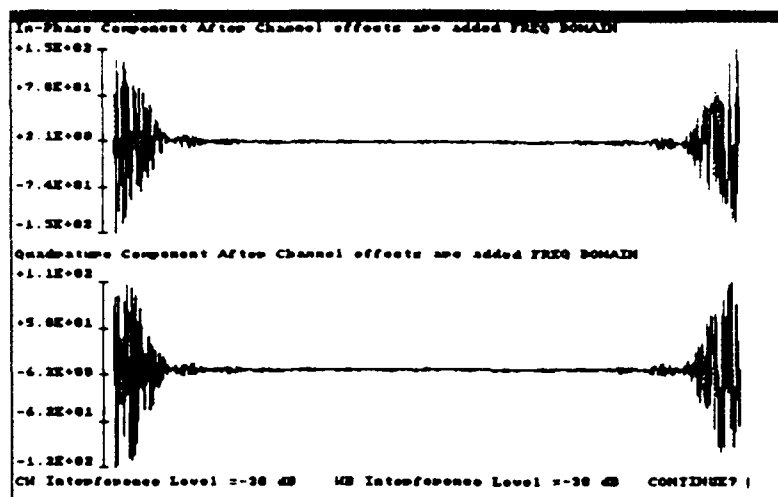


Figure 5-2

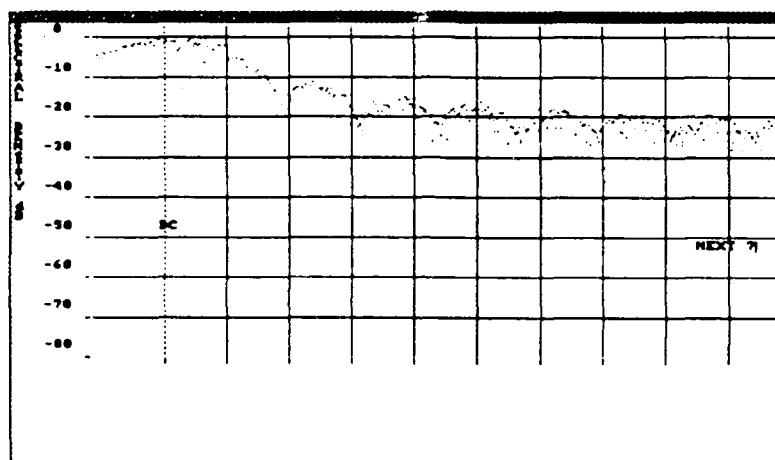


Figure 5-3

Various types of interference are added to this message by the simulator. A major consideration is large interference, usually greater than 20dB above the signal power. To gain a better perspective of how the interference affects the message it is helpful to look at the resultant signal when the interference and the message have approximately the same power. Figures (5-4) and (5-5) show the time and frequency domain representations, respectively, for a signal environment with this type of CW interference.

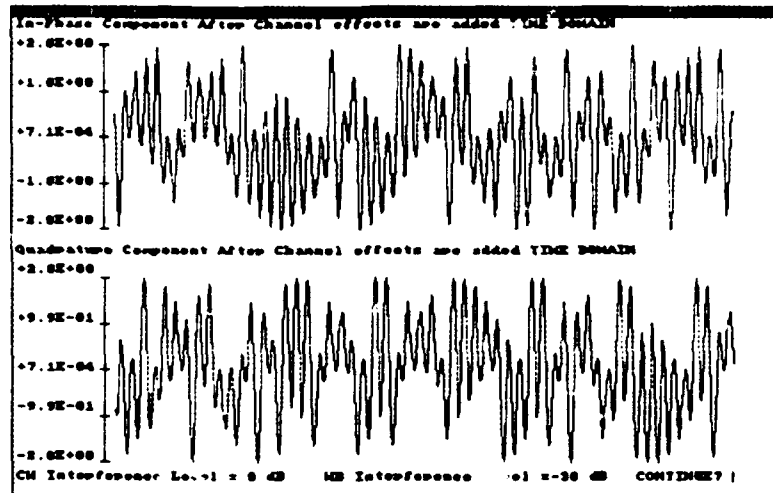


Figure 5-4

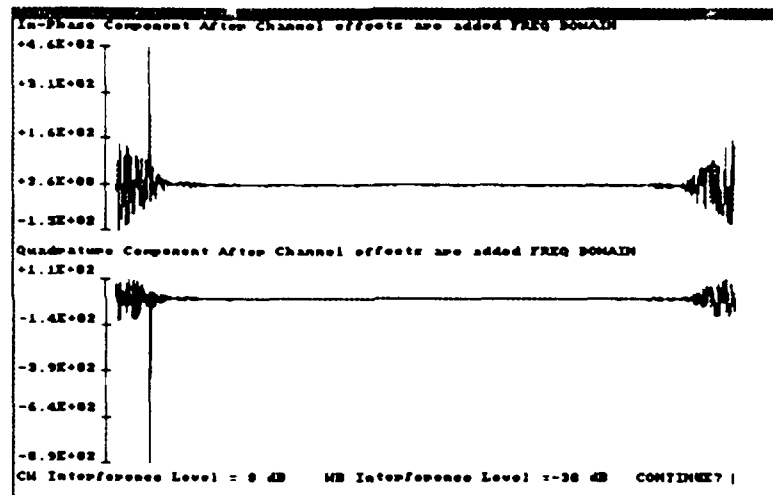


Figure 5-5

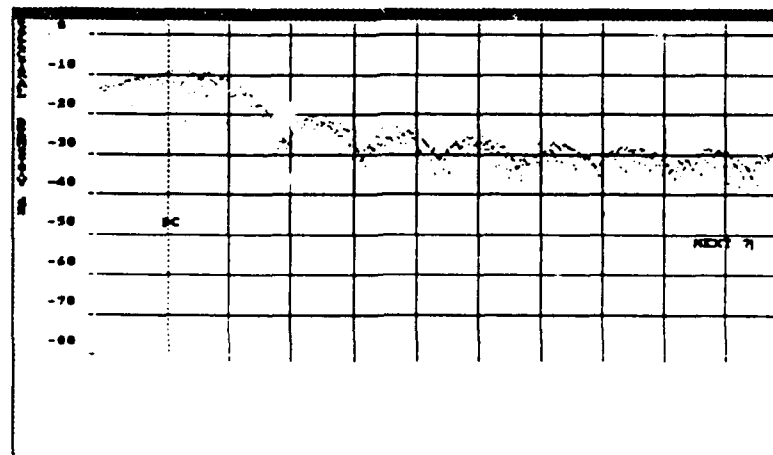


Figure 5-6

Figure (5-6) is the corresponding scatter plot. Note that the CW spike appears in Fig. (5-5), but not in Fig. (5-6). This is because of the reduced bandwidth displayed in the scatter plot. Since the CW spike determines the normalization factor, the entire plot is shifted down 10dB from the comparable plot in Fig. (5-3).

Figures (5-7) and (5-8) show the time and frequency domain representation for WB interference that is the same power as the message, and Fig. (5-9) is the scatter plot of the baseband. The upper trace on the scatter plot is the signal plus interference, and the lower trace is the reference message with no interference added.

It is the objective of this research to determine methods to recover the message from the corrupted signal. The preliminary results of this simulation are encouraging and provide an impetus for continuing research in this area. For a performance metric the present simulation uses number of bits in error out of 128 message bits. For debugging purposes and preliminary measurements this is a convenient method for making general comparisons. A better performance metric would be Bit Error Rate (BER), based on a much larger number of bits. However, a BER test would require significant modifications to the ANCP simulation and would take a substantial amount of time to compute. This issue is best addressed as an aspect of future research.

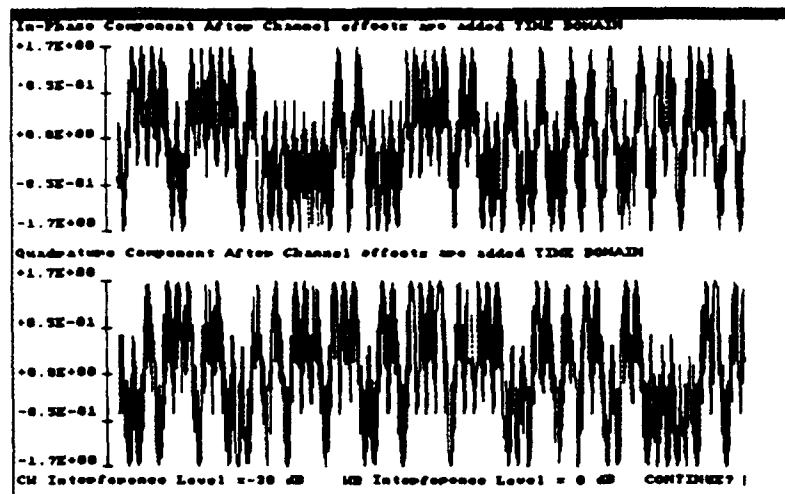


Figure 5-7

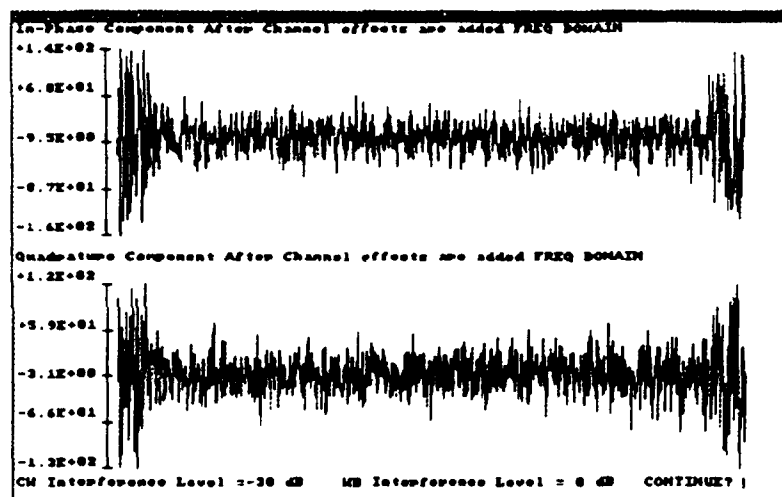


Figure 5-8

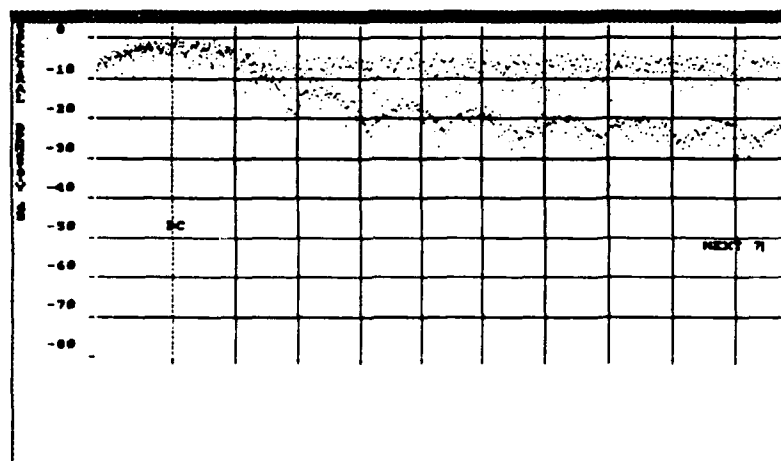


Figure 5-9

The following three examples are indicative of the capabilities of the ANCP simulation. The simulation was performed with MIPAs of order 0, 2, and 4 for different interference scenarios. More tests need to be performed before concrete conclusions can be formulated.

The first example to be considered involves the MIPA implementation of the LO demodulator with the following parameters:

CW Interference = 30dB
WB Interference = -30dB
Polynomial order = 0
Number of Intervals = 8

The simulation first adds the interference to the message. Figures (5-10) and (5-11) show the time and frequency domain representations of message plus interference, respectively. A comparison of Fig. (5-10) with Fig. (5-1) shows how much larger the interference is than the message. The simulator

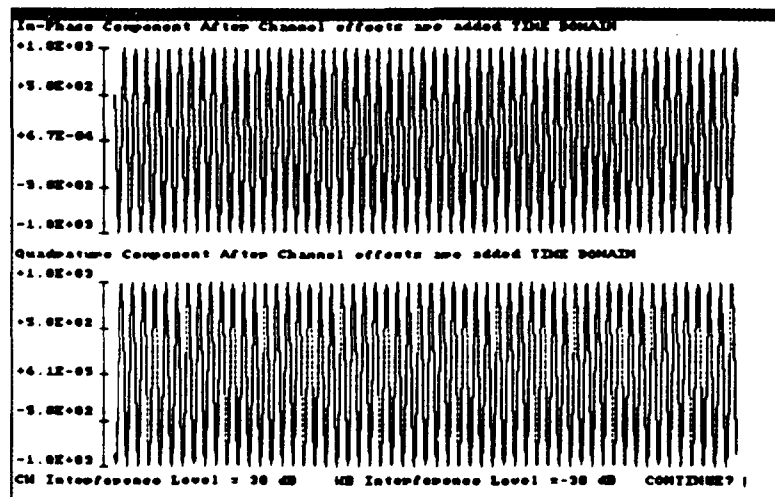


Figure 5-10

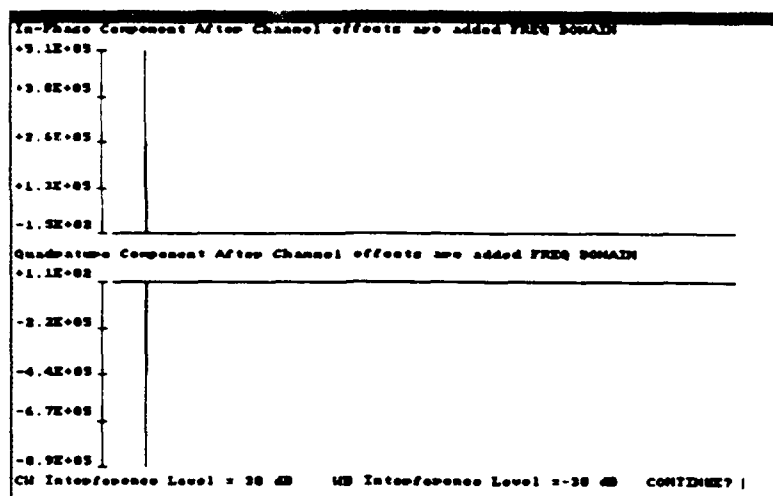


Figure 5-11

approximates the PDF of the message plus interference as shown in Fig. (5-12). From the PDF the Memoryless Nonlinear Transform (MNT) is computed, as shown in Fig. (5-13). The MNT is applied to the signal and the original message is extracted. The current ANCP simulation plots the first 34 bits of the In-Phase and Quadrature decoded messages, as shown in Figs. (5-14) and (5-15) respectively. The simulator first plots the decoded message and then superimposes the original message over it. Since only one waveform is seen in Figs. (5-14) and (5-15), the decoded message matches the original perfectly for the first 34 bits.

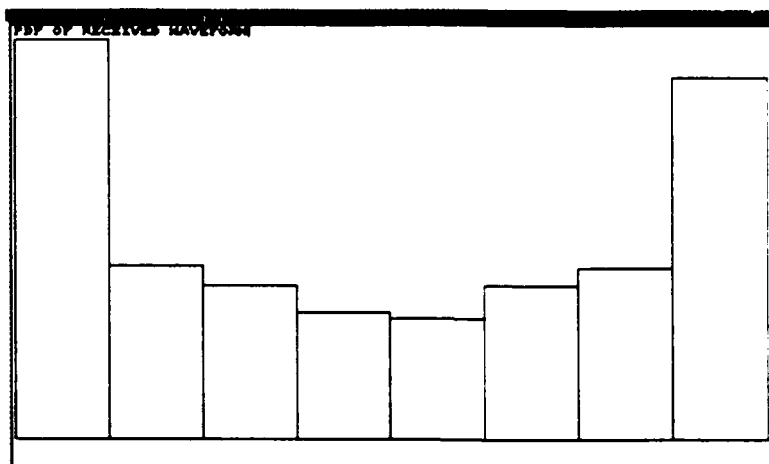


Figure 5-12

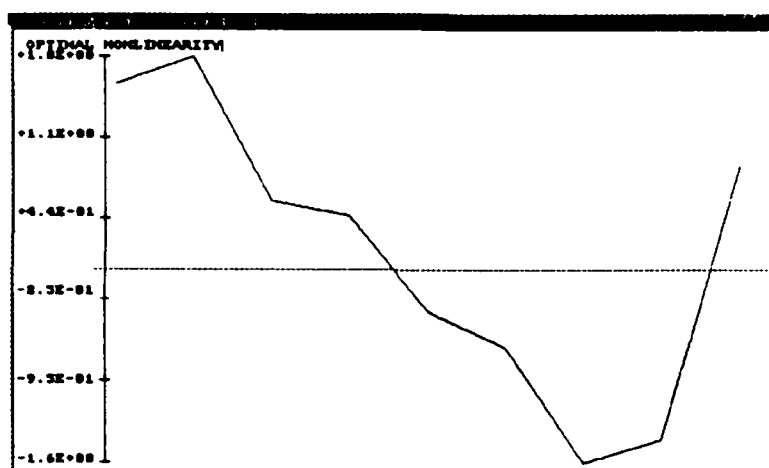


Figure 5-13

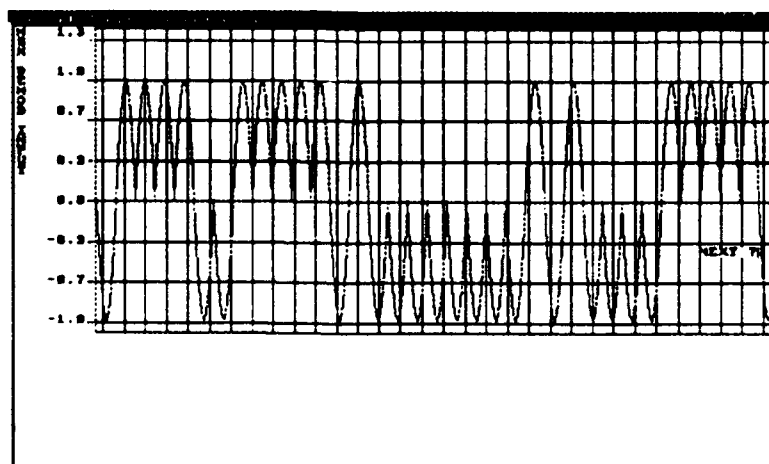


Figure 5-14

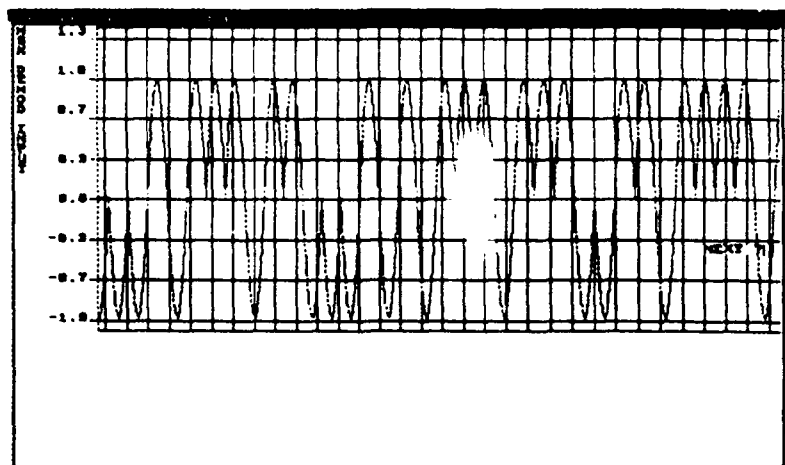


Figure 5-15

Another interesting example is:

CW Interference = -30 dB

WB Interference = 30 dB

Polynomial order = 2

Number of intervals = 2

The time domain of the message plus interference is shown in Fig. (5-16), while Fig. (5-17) depicts the corresponding frequency

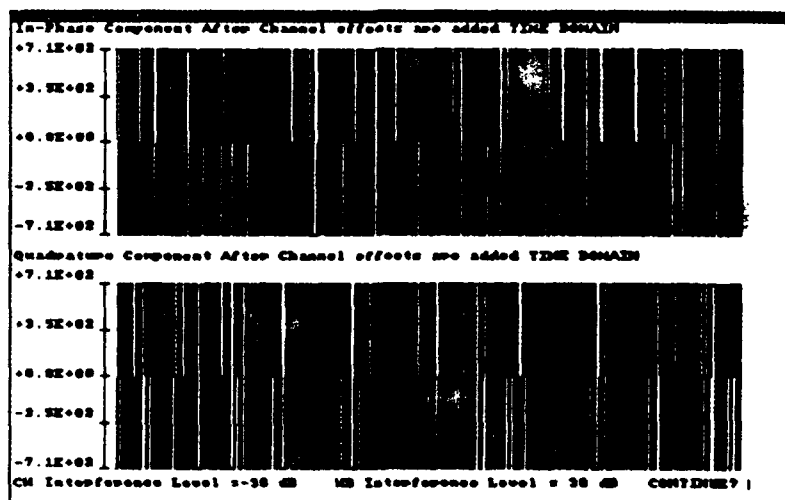


Figure 5-16

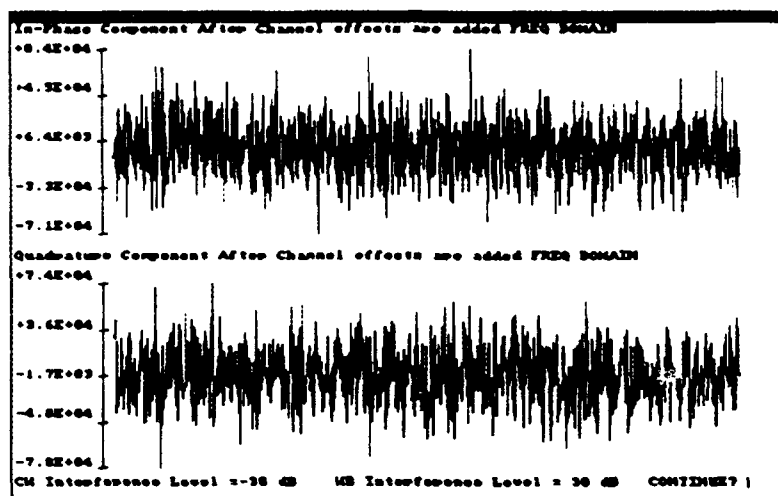


Figure 5-17

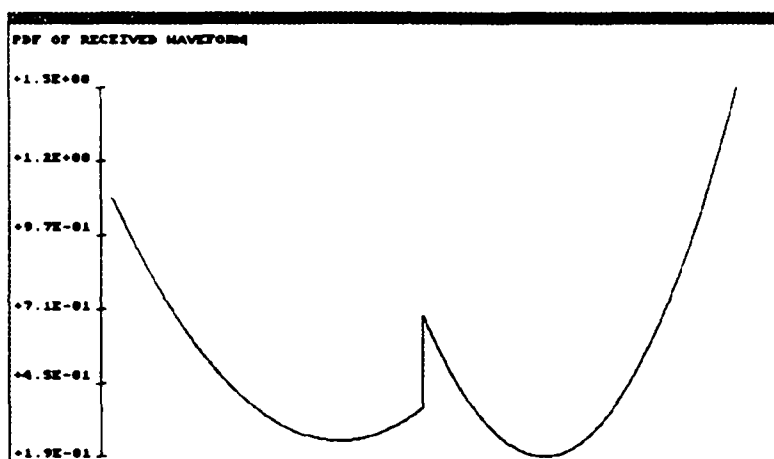


Figure 5-18

domain. The interesting phenomena in the corresponding PDF of Fig. (5-18) is the sharp discontinuity where the two intervals meet. This results because the PDF approximation is less accurate at the edges of the intervals, so the two least accurate portions of the approximation are set directly next to each other. Since the MNT is computed via a derivative, a discontinuity is seen in the resulting MNT, shown in Fig. (5-19).

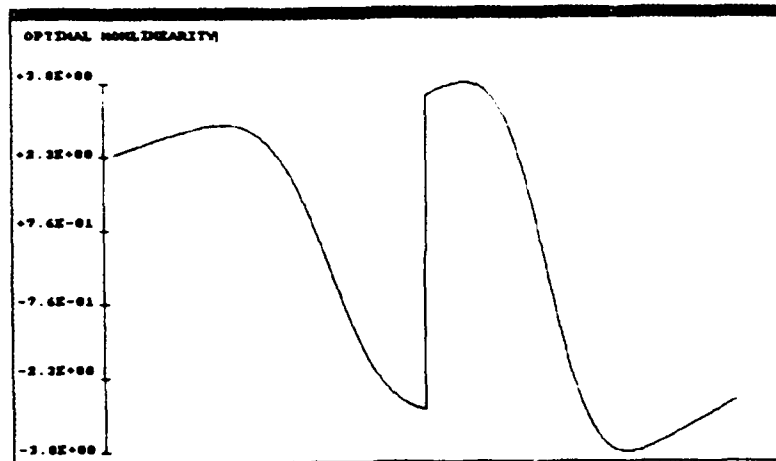


Figure 5-19

It may be possible to remedy this problem by smoothing the MIPA to the PDF. Examination of different smoothing methods is an objective of future research.

Even with the marked discontinuity of the MNT, the final result is quite close to the original message. This phenomena requires further examination. In Fig. (5-20) and (5-21) there are only three errors. Because of the nature of the MSK demodulation, the effects of these errors are minimized. The result is that only two bits out of 128 are in error at the output of the simulation.

There may be a way to avoid some of the errors that appear in Figs. (5-20) and (5-21). Notice that in all three error bits a correct decision is made for half of the bit, while an incorrect decision is made for the remaining half. This arises from the fact that the quadrature channel lags the in-phase channel by 90° . Any time a split bit occurs, it is certainly an error. An area of future research is to examine alternatives for deciding which value to choose if a split bit arises.

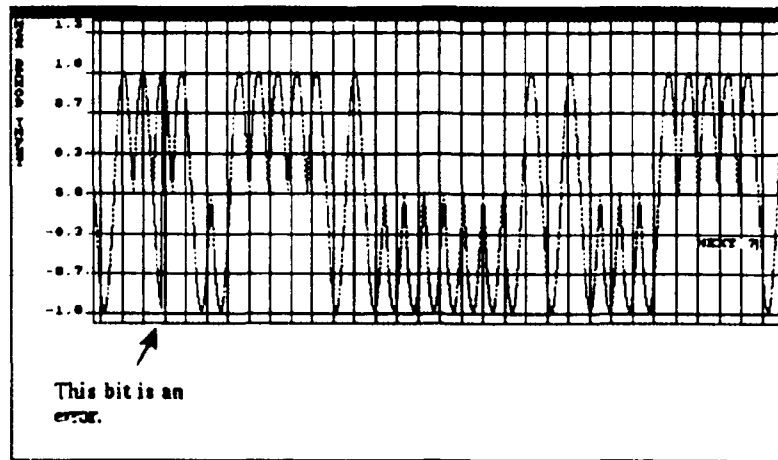


Figure 5-20

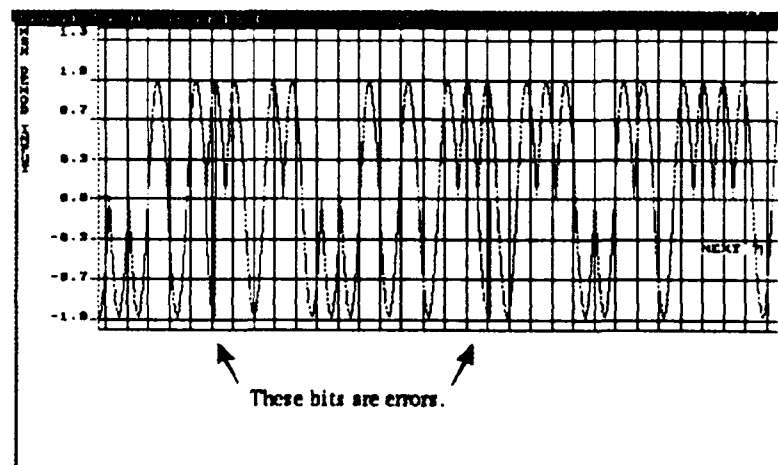


Figure 5-21

Another potential method for avoiding the discontinuity problem of Fig. (5-19) is to use just one interval to approximate the entire PDF. The problem with this method is it lacks the flexible response of the multiple interval approximations. Added flexibility can be achieved by increasing the polynomial order, but this results in more complex calculations and the need for more precision.

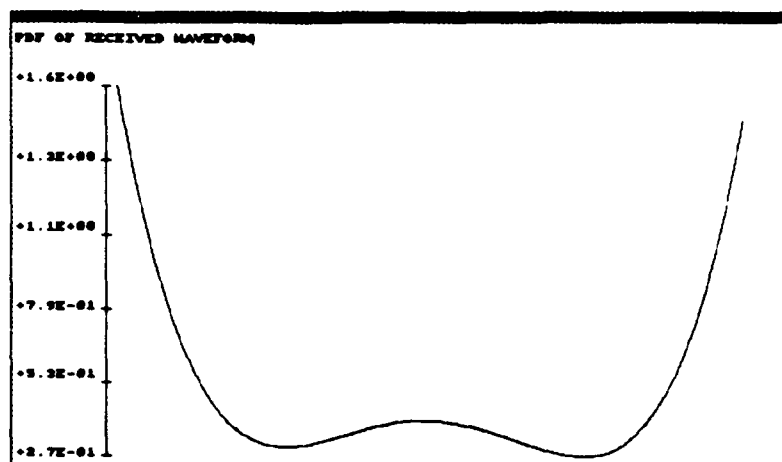


Figure 5-22

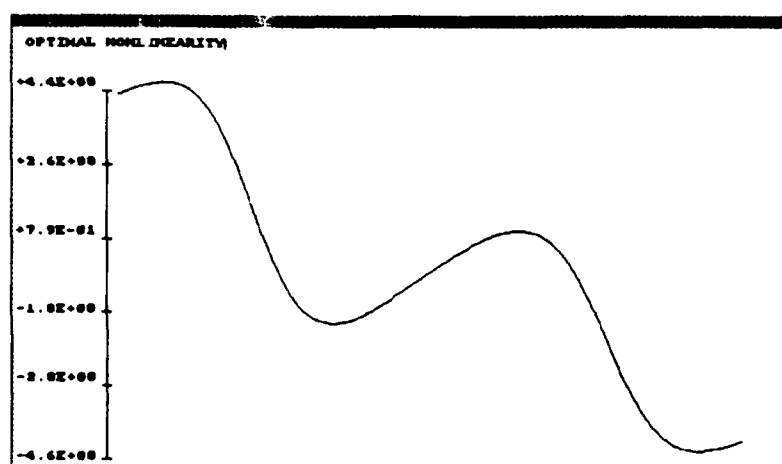


Figure 5-23

An example of a fourth order MIPA to a PDF with only one interval is shown in Fig. (5-22), and the resulting MNT is shown in Fig. (5-23). These plots are in response to 15dB of WB interference. There were no errors at the output of the simulator. Until an alternate algorithm is found or a more powerful simulator is used, the fourth order MIPA simulation can only process interference levels below 30 dB.

Both the MIPA implementation of the Locally Optimal (LO) demodulator and the Higbie approach are incorporated into the ANCP simulation. A wide range of CW and WB interference levels were applied to various orders of polynomial and number of intervals. The interference level ranges from 0dB to the limits of the simulator. At high interference levels (60dB for order 0 and 2, and 22dB for order 4) the simulator does not have enough precision to compute the results correctly. Near and below zero dB the large J/S assumption of the LO demodulator does not hold, and results are poor.

Figure (5-24) plots the simulated results for CW interference ranging from 0 to 60dB. The simulation was performed for a zero order MIPA with 8 and 16 intervals, a second order MIPA with 4 and 8 bins, and the Higbie method with 8 bins.

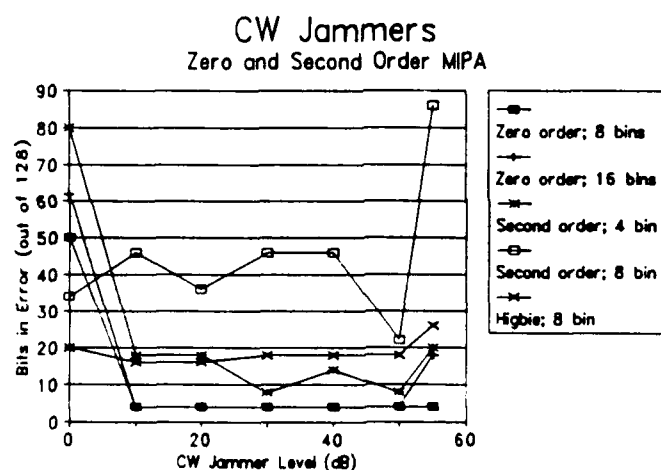


Figure 5-24

It appears that the ANCP simulation performs better when the number of intervals is small. This is mainly because there is a limited amount of samples. When the PDF is divided into many intervals, there are fewer samples in each interval. The calculation of the moments is probabalistic in nature and thus requires a large number of samples to yield adequate results. Thus, if the number of intervals is large, the MIPA will be based on inadequate moments and will produce unreliable results. This problem can be remedied by generating more samples.

Figure (5-25) shows the result for CW interference combatted by a fourth order MIPA, both with two intervals and four intervals. This results of this configuration of the ANCP simulator are not as good as described above. This is a result of several factors, including limited precision of the simulator, limited flexibility of the MIPA response, and lack of sufficient samples to achieve convergence.

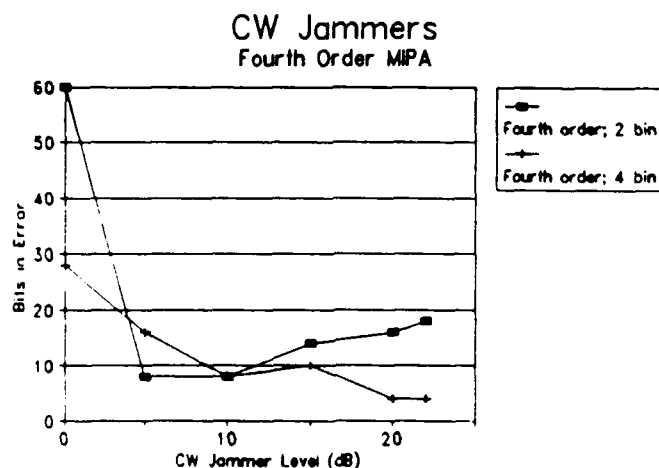


Figure 5-25

Figures (5-26) and (5-27) report the results of WB interference in the simulation. The parameters are analogous to the CW case and produce similar results.

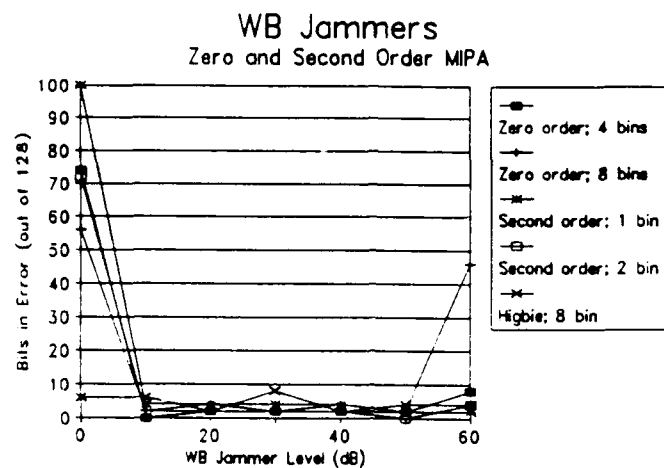


Figure 5-26

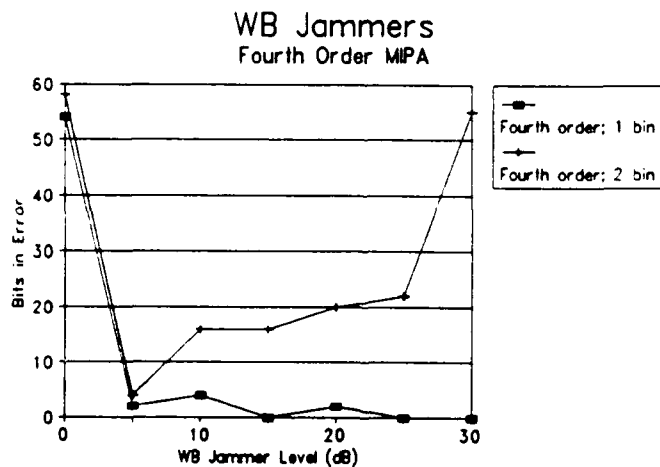


Figure 5-27

6. SUMMARY AND CONCLUSIONS

The major focus of the current research was the implementation of the MIPA nonlinear processor in the current ANCP simulation. A preliminary examination of this implementation was performed and a comparison to the nonlinear processor based on the Higbie approach was made. Comparisons were made using the number of received bits in error as a benchmark.

Derivation of the Locally Optimal (LO) maximum likelihood function was required because of an error in the Hazeltine report. The correct LO likelihood function is given by Eq. (1-4).

Solution of the transition matrix, Q , that is used for transforming normalized MIPA interval coefficients to the MIPA coefficients for the conditional PDF approximation was required. A discrepancy in the Hazeltine report made this development necessary. The correct transition matrix is given by Eq. (3-28).

The first comparison of the MIPA implementation of the LO maximum likelihood function and the Higbie approach was theoretical. It is shown in Appendix B that the optimum nonlinearities corresponding to each are equivalent. However, each method utilizes this nonlinearity differently.

Once the theoretical foundation was verified, preliminary data analysis using various scenarios was performed on each simulator implementation. From these preliminary results it appears that the performance curves, based on the number of received bits in error, for most tested simulator configurations have similar properties. At a low J/S ratio errors increase since the large interference assumption is not valid. As the J/S ratio increases, the performance of each method is better, i.e. the number of bits in error decreases. However, errors again increase as the J/S ratio increases beyond the range of the simulator. Loss of significant bits, mathematical roundoff, and arithmetic overflow are some of the causes. This indicates that a simulator with increased power and flexibility will be necessary in the future.

Preliminary data for the addition of specific types of interference was gathered. Since the results are preliminary, global conclusions should not be made at this time. More precise data must be gathered before stronger claims can be presented. For example, when a CW interferer is added, the preliminary results seem to indicate that a histogram implementation has the best performance. However, as the number of input samples increase these results will probably change. Also, if a time-varying environment is approximated, this would affect the results of each simulator configuration. For the case of WB noise, it is not possible to differentiate which configuration performs best from the preliminary data.

It will be necessary to examine the nature and relevance of the current WB noise model. Most importantly, the preliminary results illustrated in Fig. (5-22) through Fig. (5-25) show that each nonlinear processing method is capable of countering the effects of non-Gaussian interference, at least under specific conditions.

One note of special interest: The preliminary results indicate that the best performance for each simulator configuration occurs when a low number of intervals, or quantiles, is selected. This is most probably a result of the low number of input samples used by the simulation. As shown in the Hazeltine report^[1], with input samples on the order of 10,000, a marked improvement can be expected when the number of intervals is increased. Currently, the ANCP simulation uses 1024 samples.

The present research and the summary presented here indicate that future efforts will have promising results. The future objectives of this research are presented in the next section.

7. FUTURE OBJECTIVES

The preliminary work on this project suggests that the MIPA and Higbie nonlinear processors have the potential to effectively mitigate the effects of non-Gaussian jammers. However, to determine which processing scheme has the best overall results in a given environment, further efforts in this area are necessary. A major goal of the future research is a more complete study of various nonlinear processing techniques. This will facilitate the determination of the optimal processor. Future objectives of the project include:

- A. Using the ANCP simulation to extensively verify the results of the Hazeltine report. Particular emphasis will be placed on correcting any discrepancies found.
- B. Determining the effect of varying the number and size of the intervals in the MIPA algorithm.
- C. Determining the order of approximating polynomial which provides optimal performance and realistic complexity.
- D. Determining if the wide band noise currently being added to the information signal is Gaussian, and if not, incorporating a Gaussian wide band noise model into the simulation.

- E. Determining the effects of adding impulsive noise and impulse excision.
- F. Determining the effects of smoothing on the MIPA algorithm.
- G. Implementing the Hazeltine memoryless nonlinear transform (MNT), i.e. the optimum nonlinear transform, directly using the MIPA algorithm.
- H. Examining the benefits of combining the MIPA algorithm and the Higbie approach into one method.
- I. Modifying the current simulation to periodically update the probability density function (PDF) for the MIPA algorithm, or the cumulative distribution function (CDF) for the Higbie approach. This will allow investigation into the area of training cycles, that is time necessary for a processor to learn the environment before actual signal processing begins. Also, examination of the time varying nature of the signal environment will be possible.
- J. Examination of another approach developed by the Charles Stark Draper Laboratory^{[6][7]}. This would include verification of the Draper reports and

implementation of the Draper Digital Density Detector in the ANCP simulation. In addition, an independent simulation would be developed for examination of the Draper Robust Digital Adaptive Transceiver, a method that integrates spatial, temporal, and amplitude processing.

The anticipated results of this future work will illustrate the tradeoffs of each nonlinear processor algorithm for use in a spread spectrum receiver. With this knowledge, the most effective processor for a given interference scenario can be determined.

**APPENDIX A: COMPARISON BETWEEN THE LOCALLY OPTIMAL AND
HIGBIE OPTIMUM NONLINEARITIES FOR TWO-
DIMENSIONAL SIGNALS**

The Higbie representation of a two-dimensional signal, after sampling, is given by,

$$I_x + jQ_x \quad (A-1)$$

where: I_x is the in-phase component,
 Q_x is the quadrature component.

or in polar form,

$$A_x e^{j\phi_x} \quad (A-2)$$

where: $A_x = \sqrt{I_x^2 + Q_x^2}$ is the magnitude,
 $\phi_x = \arctan \frac{Q_x}{I_x}$ is the phase.

The corresponding Higbie optimum nonlinearity can be written as,

$$y = \{A_y^I + jA_y^Q\} e^{j\phi_x} \quad (A-3)$$

with,

$$A_y^I = -\left(\frac{A_s^2}{2}\right) \frac{\frac{\partial}{\partial A_x} \left\{ \frac{P_{A\phi}(A_x, \phi_x)}{A_x} \right\}}{\left\{ \frac{P_{A\phi}(A_x, \phi_x)}{A_x} \right\}} \quad (\text{A-4})$$

and,

$$A_y^Q = -\left(\frac{A_s^2}{2}\right) \left(\frac{1}{A_x}\right) \frac{\frac{\partial}{\partial \phi_x} P_{A\phi}(A_x, \phi_x)}{P_{A\phi}(A_x, \phi_x)} \quad (\text{A-5})$$

where: $P_{A\phi}(A_x, \phi_x)$ is the received signal probability density function (PDF) in polar form,
 A_s is a constant.

To show the equivalence between the Higbie and the Locally Optimal nonlinearity, it is necessary to derive the Higbie nonlinearity for the received signal PDF in rectangular form. This PDF can be derived from the polar form PDF using the relation,

$$P_{A\phi}(A_x, \phi_x) = A_x \times P_{IQ}(i, q) \quad (\text{A-6})$$

where: $i = i(A_x, \phi_x) = A_x \times \cos \phi_x$ and $q = q(A_x, \phi_x) = A_x \times \sin \phi_x$.

Substituting Eq. (A-6) into Eq. (A-4), the result is,

$$A_y^I = -\frac{A_s^2}{2} \frac{\frac{\partial}{\partial A_x} p_{I0}(i, q)}{p_{I0}(i, q)} \quad (\text{A-7})$$

Note that,

$$\frac{\partial}{\partial A_x} p_{I0}(i, q) = \frac{\partial}{\partial i} p_{I0}(i, q) \frac{\partial i}{\partial A_x} + \frac{\partial}{\partial q} p_{I0}(i, q) \frac{\partial q}{\partial A_x} \quad (\text{A-8})$$

with $\frac{\partial i}{\partial A_x} = \cos \phi_x$ and $\frac{\partial q}{\partial A_x} = \sin \phi_x$.

Therefore,

$$A_y^I = -\frac{A_s^2}{2} \left[\cos \phi_x \times \frac{\frac{\partial}{\partial i} p_{I0}(i, q)}{p_{I0}(i, q)} + \sin \phi_x \times \frac{\frac{\partial}{\partial q} p_{I0}(i, q)}{p_{I0}(i, q)} \right] \quad (\text{A-9})$$

Similarly, substituting Eq. (A-6) into Eq. (A-5), the result is,

$$A_y^O = -\frac{A_s^2}{2} \left(\frac{1}{A_x} \right) \frac{\frac{\partial}{\partial \phi_x} A_x \times p_{I0}(i, q)}{A_x \times p_{I0}(i, q)} \quad (\text{A-10})$$

Similar to the development of Eq. (A-9), note that,

$$\frac{\partial}{\partial \phi_x} p_{I0}(i, q) = \frac{\partial}{\partial i} p_{I0}(i, q) \frac{\partial i}{\partial \phi_x} + \frac{\partial}{\partial q} p_{I0}(i, q) \frac{\partial q}{\partial \phi_x} \quad (\text{A-11})$$

with $\frac{\partial i}{\partial \phi_x} = -A_x \sin \phi_x$ and $\frac{\partial q}{\partial \phi_x} = A_x \cos \phi_x$.

Therefore,

$$A_y^Q = -\frac{A_s^2}{2} \left[-(\sin\phi_x) \times \frac{\frac{\partial}{\partial i} P_{IQ}(i, q)}{P_{IQ}(i, q)} + \cos\phi_x \times \frac{\frac{\partial}{\partial q} P_{IQ}(i, q)}{P_{IQ}(i, q)} \right] \quad (\text{A-12})$$

Now, with A_y^I and A_y^Q in rectangular coordinates it is possible to substitute Eq. (A-9) and Eq. (A-12) into the Eq. (A-3). Noticing that $e^{j\phi} = \cos\phi + j\sin\phi$, the optimum nonlinearity becomes,

$$y = \{A_y^I + jA_y^Q\} e^{j\phi_x} = \{A_y^I + jA_y^Q\} \{\cos\phi_x + j\sin\phi_x\} \quad (\text{A-13})$$

or,

$$y = \{A_y^I \cos\phi_x - A_y^Q \sin\phi_x\} + j\{A_y^I \sin\phi_x + A_y^Q \cos\phi_x\} \quad (\text{A-14})$$

Now, calculate the real part of y:

$$\text{Re}(y) = A_y^I \cos\phi_x - A_y^Q \sin\phi_x =$$

$$-\frac{A_s^2}{2} \cos\phi_x \left[\cos\phi_x \times \frac{\frac{\partial}{\partial i} P_{IQ}(i, q)}{P_{IQ}(i, q)} + \sin\phi_x \times \frac{\frac{\partial}{\partial q} P_{IQ}(i, q)}{P_{IQ}(i, q)} \right] \quad (\text{A-15})$$

$$+ \frac{A_s^2}{2} \sin\phi_x \left[(-\sin\phi_x) \frac{\frac{\partial}{\partial i} P_{IQ}(i, q)}{P_{IQ}(i, q)} + \cos\phi_x \times \frac{\frac{\partial}{\partial q} P_{IQ}(i, q)}{P_{IQ}(i, q)} \right]$$

$$Re(y) = -\frac{A_s^2}{2} \left[(\cos^2 \phi_x + \sin^2 \phi_x) \times \frac{\frac{\partial}{\partial i} P_{I0}(i, q)}{P_{I0}(i, q)} \right]$$

(A-16)

$$-\frac{A_s^2}{2} \left[(\cos \phi_x \sin \phi_x - \cos \phi_x \sin \phi_x) \times \frac{\frac{\partial}{\partial q} P_{I0}(i, q)}{P_{I0}(i, q)} \right]$$

Since $\cos^2 \phi_x + \sin^2 \phi_x = 1$:

$$Re(y) = -\frac{A_s^2}{2} \frac{\frac{\partial}{\partial i} P_{I0}(i, q)}{P_{I0}(i, q)} \quad (A-17)$$

Similarly, it is possible to calculate the imaginary part of y:

$$Im(y) = A_y^I \sin \phi_x + A_y^Q \cos \phi_x =$$

$$-\frac{A_s^2}{2} \sin \phi_x \left[\cos \phi_x \times \frac{\frac{\partial}{\partial i} P_{I0}(i, q)}{P_{I0}(i, q)} + \sin \phi_x \times \frac{\frac{\partial}{\partial q} P_{I0}(i, q)}{P_{I0}(i, q)} \right] \quad (A-18)$$

$$-\frac{A_s^2}{2} \cos \phi_x \left[(-\sin \phi_x) \frac{\frac{\partial}{\partial i} P_{I0}(i, q)}{P_{I0}(i, q)} + \cos \phi_x \times \frac{\frac{\partial}{\partial q} P_{I0}(i, q)}{P_{I0}(i, q)} \right]$$

$$\text{Im}(y) = -\frac{A_s^2}{2} \left[(\cos\phi_x \sin\phi_x - \cos\phi_x \sin\phi_x) \times \frac{\frac{\partial}{\partial i} p_{IQ}(i, q)}{p_{IQ}(i, q)} \right] \quad (\text{A-19})$$

$$-\frac{A_s^2}{2} \left[(\cos^2\phi_x + \sin^2\phi_x) \times \frac{\frac{\partial}{\partial q} p_{IQ}(i, q)}{p_{IQ}(i, q)} \right]$$

Using the previous trigonometric identity:

$$\text{Im}(y) = -\frac{A_s^2}{2} \frac{\frac{\partial}{\partial q} p_{IQ}(i, q)}{p_{IQ}(i, q)} \quad (\text{A-20})$$

Therefore, the resulting Higbie optimum nonlinearity is:

$$y = \left[-\frac{A_s^2}{2} \frac{\frac{\partial}{\partial i} p_{IQ}(i, q)}{p_{IQ}(i, q)} \right] + j \left[-\frac{A_s^2}{2} \frac{\frac{\partial}{\partial q} p_{IQ}(i, q)}{p_{IQ}(i, q)} \right] \quad (\text{A-21})$$

Compare this to the Locally Optimal maximum likelihood function, given by:

$$\bar{L}_j = -\sum_{k=1}^K \left\{ r_{kj} \times \frac{\frac{\partial}{\partial i} f_{IQ}(i_k, q_k)}{f_{IQ}(i_k, q_k)} + s_{kj} \times \frac{\frac{\partial}{\partial q} f_{IQ}(i_k, q_k)}{f_{IQ}(i_k, q_k)} \right\} \quad (3-6)$$

It can be seen that the in-phase and quadrature optimum nonlinearities are identical, i.e.:

$$\frac{\frac{\partial}{\partial i} p_{I0}(i, q)}{p_{I0}(i, q)} = \frac{\frac{\partial}{\partial i} f_{I0}(i, q)}{f_{I0}(i, q)} \quad (\text{A-22})$$

$$\frac{\frac{\partial}{\partial q} p_{I0}(i, q)}{p_{I0}(i, q)} = \frac{\frac{\partial}{\partial q} f_{I0}(i, q)}{f_{I0}(i, q)} \quad (\text{A-23})$$

However, each method implements these optimum nonlinearities differently.

Next, examine the case when the received signal PDF is assumed to have bivariate radial symmetry. The received PDF becomes:

$$p_{A\phi}(A_x, \phi_x) = \begin{cases} \frac{p_A(A_x)}{2\pi}, & 0 < \phi < 2\pi \\ 0, & \text{elsewhere} \end{cases} \quad (\text{A-24})$$

Then A_y^I becomes:

$$A_y^I = -\frac{A_s^2}{2} \frac{\frac{d}{dA_x} \left\{ \frac{p_A(A_x)}{2\pi A_x} \right\}}{\frac{p_A(A_x)}{2\pi A_x}} = -\frac{A_s^2}{2} \left[\frac{A_x}{p_A(A_x)} \right] \left[-\frac{p_A(A_x)}{A_x^2} + \frac{\frac{d}{dA_x} p_A(A_x)}{A_x} \right] \quad (\text{A-25})$$

Multiplying and rearranging terms:

$$A_y^I = -\frac{A_s^2}{2} \left[\frac{\frac{d}{dA_x} p_A(A_x)}{p_A(A_x)} - \frac{1}{A_x} \right] \quad (\text{A-26})$$

Next, examine A_y^Q :

$$A_y^Q = -\frac{A_s^2}{2} \left(\frac{1}{A_x} \right) \frac{\frac{\partial}{\partial \phi_x} \frac{p_A(A_x)}{2\pi}}{\frac{p_A(A_x)}{2\pi}} = 0 \quad (\text{A-27})$$

Therefore, the Higbie nonlinearity of Eq. (A-3) reduces to:

$$y = A_y^I e^{j\phi_x} = (\cos\phi_x) A_y^I + j(\sin\phi_x) A_y^I \quad (\text{A-28})$$

$$y = \frac{A_s^2}{2} \left\{ \cos\phi_x \times \left[-\frac{\frac{d}{dA_x} p_A(A_x)}{p_A(A_x)} + \frac{1}{A_x} \right] + j \sin\phi_x \times \left[-\frac{\frac{d}{dA_x} p_A(A_x)}{p_A(A_x)} + \frac{1}{A_x} \right] \right\} \quad (\text{A-29})$$

Compare this to the Locally Optimal maximum likelihood function, given by:

$$\bar{L}_j = \sum_{k=1}^K \left\{ r_{kj} \times \cos\theta \times \left[-\frac{\frac{d}{dr} f_e(r)}{f_e(r)} + \frac{1}{r} \right] + s_{kj} \times \sin\theta \times \left[-\frac{\frac{d}{dr} f_e(r)}{f_e(r)} + \frac{1}{r} \right] \right\} \quad (\text{1-5})$$

It can again be seen that the in-phase and quadrature optimum nonlinearities are identical, i.e.:

$$-\frac{\frac{d}{dA_x} p_A(A_x)}{p_A(A_x)} + \frac{1}{A_x} = -\frac{\frac{d}{dr} f_\theta(r)}{f_\theta(r)} + \frac{1}{r} \quad (\text{A-30})$$

As stated earlier, each method implements these nonlinearities differently. The result is two different nonlinear processors based on the same optimum nonlinearity.

APPENDIX B: DERIVATION OF THE LOCALLY OPTIMAL MAXIMUM LIKELIHOOD FUNCTION FOR COMPLEX SIGNALS

The transmitted, or source, signal is represented by a random process with an in-phase component $R(t)$ and a quadrature component $S(t)$. For the received signal, the in-phase component is $I(t)$ and the quadrature component is $Q(t)$. Finally, the interference is characterized by an in-phase signal $X(t)$, and a quadrature signal $Y(t)$. After sampling at the receiver input, the following relation results:

$$I=R+X \quad Q=S+Y \quad (B-1)$$

The cumulative distribution function, $F_{IQ}(i,q)$, is calculated using the above relations, i.e.:

$$F_{IQ}(i,q)=P(I\leq i, Q\leq q)=P(R+X\leq i, S+Y\leq q) \quad (B-2)$$

Therefore:

$$F_{IQ}(i,q)=\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{q-s} \int_{-\infty}^{i-r} f_{RSXY}(r,s,x,y) dx dy dr ds \quad (B-3)$$

where: $f_{RSXY}(r,s,x,y)$ is the joint PDF of the random variables r,s,x , and y .

If the transmitted signal is independent of the noise, then:

$$F_{IQ}(i, q) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{q-s} \int_{-\infty}^{i-r} f_{RS}(r, s) f_{XY}(x, y) dx dy dr ds \quad (B-4)$$

where: $f_{RS}(r, s)$ is the joint PDF of the transmitted signal

$f_{XY}(x, y)$ is the joint PDF of the interference.

Let $H_{XY}(x, y) = \iint f_{XY}(x, y) dx dy$. Substituting:

$$F_{IQ}(i, q) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{RS}(r, s) \{ [H_{XY}(i-r, q-s) - H_{XY}(i-r, -\infty)] - [H_{XY}(-\infty, q-s) - H_{XY}(-\infty, -\infty)] \} dr ds \quad (B-5)$$

The PDF of the received signal, $f_{IQ}(i, q)$, is obtained by differentiation:

$$f_{IQ}(i, q) = \frac{\partial^2}{\partial i \partial q} F_{IQ}(i, q) \quad (B-6)$$

$$f_{IQ}(i, q) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{RS}(r, s) \frac{\partial^2}{\partial i \partial q} \{ [H_{XY}(i-r, q-s) - H_{XY}(i-r, -\infty)] - [H_{XY}(-\infty, q-s) - H_{XY}(-\infty, -\infty)] \} dr ds \quad (B-7)$$

$$f_{IQ}(i, q) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{RS}(r, s) \{f_{XY}(i-r, q-s) - 0-0+0\} dr ds \quad (B-8)$$

Therefore:

$$f_{IQ}(i, q) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{RS}(r, s) f_{XY}(i-r, q-s) dr ds \quad (B-9)$$

Next, if the transmitted signal is one of N equiprobable transmitted signal sets, then:

$$f_{RS}(r, s) = \frac{1}{N} \sum_{j=1}^N \delta(r-r_j) \delta(s-s_j) \quad (B-10)$$

where: r_j and s_j are one of N possible signal sets,
 $\delta(\cdot)$ is the discrete delta function.

The resulting received signal PDF is:

$$f_{IQ}(i, q) = \frac{1}{N} \sum_{j=1}^N \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(i-r_j, q-s_j) \delta(r-r_j) \delta(s-s_j) dr ds \quad (B-11)$$

or:

$$f_{IQ}(i, q) = \frac{1}{N} \sum_{j=1}^N f_{XY}(i-r_j, q-s_j) \quad (B-12)$$

For maximum likelihood detection, it is desired to maximize $P(R=r_j, S=s_j | I=i, Q=q)$. This is equivalent to choosing r_j and s_j , $j=1 \rightarrow N$, to maximize $f_{XY}(i-r_j, q-s_j)$. However, if we assume that K samples of each received signal are used to determine the maximum probability, then the transmitted signal set is increased to an N -ary signal set in which each signal is a vector of length K . This can be written as $\vec{I} = \vec{R} + \vec{X}$ and $\vec{Q} = \vec{R} + \vec{Y}$. Assuming independent and identically distributed samples, then the received signal PDF is given by:

$$f_{\vec{I}\vec{Q}}(\vec{I}, \vec{Q}) = \frac{1}{N} \sum_{j=1}^N \left\{ \prod_{k=1}^K f_{XY}(i_k - r_{jk}, q_k - s_{jk}) \right\} \quad (\text{B-13})$$

Therefore, for maximum likelihood detection, choose \vec{r}_j and \vec{s}_j , $j=1 \rightarrow N$, to maximize:

$$L_j' = \prod_{k=1}^K f_{XY}(i_k - r_{jk}, q_k - s_{jk}) \quad (\text{B-14})$$

where: K is the length of a signal vector.

To simplify calculations, the natural logarithm of this function is used. This results in the Globally Optimal (GO) maximum likelihood function:

Choose \bar{r}_j and \bar{s}_j , $j=1 \rightarrow N$, to maximize

$$L_j = \sum_{k=1}^K \ln[f_{XY}(i_k - r_{jk}, q_k - s_{jk})] \quad (\text{B-15})$$

where: i_k, q_k is the k^{th} sample of the received signal
for the I and Q channels respectively,
 r_{kj}, s_{kj} is the k^{th} of K samples used to
represent the j^{th} transmitted signal pair.

To further simplify calculations, a Taylor Series approximation of the interference PDF is used. The expansion is around the point (i_k, q_k) since this value is known. This approximation is useful if a large interference to signal ratio is assumed. Therefore:

$$\begin{aligned} f_{XY}(i_k - r_{jk}, q_k - s_{jk}) \approx & f_{XY}(i_k, q_k) + \frac{\partial}{\partial X} f_{XY}(i_k, q_k) \times (x_k - i_k) \\ & + \frac{\partial}{\partial Y} f_{XY}(i_k, q_k) \times (y_k - i_k) \end{aligned} \quad (\text{B-16})$$

Recalling that $x_k = i_k - r_{jk}$, $y_k = q_k - s_{jk}$:

$$f_{XY}(i_k - r_{jk}, q_k - s_{jk}) \approx f_{XY}(i_k, q_k) - r_{jk} \times \frac{\partial}{\partial X} f_{XY}(i_k, q_k) - s_{jk} \times \frac{\partial}{\partial Y} f_{XY}(i_k, q_k) \quad (\text{B-17})$$

Substituting this result into the GO maximum likelihood function:

$$L_j \approx \sum_{k=1}^K \ln \left[f_{XY}(i_k, q_k) - r_{jk} \times \frac{\partial}{\partial X} f_{XY}(i_k, q_k) - s_{jk} \times \frac{\partial}{\partial Y} f_{XY}(i_k, q_k) \right] \quad (\text{B-18})$$

$$L_j \approx \sum_{k=1}^K \left\{ \ln[f_{XY}(i_k, q_k)] + \ln \left[1 - r_{jk} \frac{\frac{\partial}{\partial X} f_{XY}(i_k, q_k)}{f_{XY}(i_k, q_k)} - s_{jk} \frac{\frac{\partial}{\partial Y} f_{XY}(i_k, q_k)}{f_{XY}(i_k, q_k)} \right] \right\} \quad (\text{B-19})$$

Since we are comparing \bar{r}_j , \bar{s}_j to maximize L_j , an equivalent likelihood function is:

$$\bar{L}_j = \sum_{k=1}^K \ln \left[1 - r_{jk} \frac{\frac{\partial}{\partial X} f_{XY}(i_k, q_k)}{f_{XY}(i_k, q_k)} - s_{jk} \frac{\frac{\partial}{\partial Y} f_{XY}(i_k, q_k)}{f_{XY}(i_k, q_k)} \right] \quad (\text{B-20})$$

Using the approximation $\ln(1-x) \approx -x$:

$$\bar{L}_j = - \sum_{k=1}^K \left\{ r_{jk} \frac{\frac{\partial}{\partial X} f_{XY}(i_k, q_k)}{f_{XY}(i_k, q_k)} + s_{jk} \frac{\frac{\partial}{\partial Y} f_{XY}(i_k, q_k)}{f_{XY}(i_k, q_k)} \right\} \quad (\text{B-21})$$

Finally, since a high interference to signal ratio was assumed, $f_{XY}(i, q) \approx f_{IQ}(i, q)$ and $R \ll X$, $S \ll Y$. This results in the Locally Optimal (LO) maximum likelihood function:

Choose \bar{r}_j and \bar{s}_j , $j=1 \rightarrow N$, to maximize

$$\bar{L}_j = - \sum_{k=1}^K \left\{ r_{kj} \times \frac{\frac{\partial}{\partial I} f_{IQ}(i_k, q_k)}{f_{IQ}(i_k, q_k)} + s_{kj} \times \frac{\frac{\partial}{\partial Q} f_{IQ}(i_k, q_k)}{f_{IQ}(i_k, q_k)} \right\} \quad (\text{B-22})$$

Further simplification of the LO maximum likelihood function results when complex radial symmetry of the received signal PDF is assumed. Under this assumption:

$$f_{IQ}(i, q) = \begin{cases} \frac{f_e(r)}{2\pi r}, & 0 < \theta < 2\pi \\ 0, & \text{elsewhere} \end{cases} \quad (\text{B-23})$$

where: $f_{IQ}(\cdot)$ is the received signal PDF,
 $f_e(r)$ is the received envelope PDF,
 $r = \sqrt{i^2 + q^2}$ is the received magnitude,
 $\theta = \arctan \frac{q}{i}$ is the received phase angle.

To derive the expression for the LO likelihood function, first the expressions for the partial derivatives must be shown:

$$\frac{\partial}{\partial I} f_{IQ}(i, q) = \frac{\partial}{\partial I} \frac{f_e(r)}{2\pi r} = \frac{d}{dr} \frac{f_e(r)}{2\pi r} \times \frac{\partial r}{\partial I} \quad (\text{B-24})$$

Since $\frac{\partial r}{\partial i} = \cos\theta$:

$$\frac{\partial}{\partial i} f_{IQ}(i, q) = \cos\theta \times \frac{d}{dr} \frac{f_e(r)}{2\pi r} \quad (\text{B-25})$$

Similarly:

$$\frac{\partial}{\partial q} f_{IQ}(i, q) = \frac{\partial}{\partial q} \frac{f_e(r)}{2\pi r} = \frac{d}{dr} \frac{f_e(r)}{2\pi r} \times \frac{\partial r}{\partial q} \quad (\text{B-26})$$

Since $\frac{\partial r}{\partial q} = \sin\theta$:

$$\frac{\partial}{\partial q} f_{IQ}(i, q) = \sin\theta \times \frac{d}{dr} \frac{f_e(r)}{2\pi r} \quad (\text{B-27})$$

Next:

$$\frac{d}{dr} \frac{f_e(r)}{2\pi r} = \frac{\frac{d}{dr} f_e(r)}{2\pi r} - \frac{f_e(r)}{2\pi r^2} \quad (\text{B-28})$$

Substituting these results into the expression for the LO maximum likelihood function of Eq. (B-22), the result is:

Choose \bar{r}_j and \bar{s}_j , $j=1 \rightarrow N$, to maximize

$$\bar{L}_j = \sum_{k=1}^K \left\{ r_{kj} \times \cos\theta \times \left[-\frac{\frac{d}{dr} f_e(r)}{f_e(r)} + \frac{1}{r} \right] + s_{kj} \times \sin\theta \times \left[-\frac{\frac{d}{dr} f_e(r)}{f_e(r)} + \frac{1}{r} \right] \right\} \quad (\text{B-29})$$

REFERENCES

- [1] Hazeltine Report No. 6662, Adaptive Nonlinear Coherent Processor Design, Vols. I and II, RADC Contract No. F30602-86-C-0106, October 18, 1988.
- [2] Higbie, J.H., "Adaptive Nonlinear Suppression of Interference", IEEE Proceedings of MILCOM '88, October 1988.
- [3] Stark, H. and Woods, J., Probability, Random Processes, and Estimation for Engineers, New Jersey: Prentice-Hall, 1986.
- [4] Pasupathy, S., "Minimum Shift Keying: A Spectrally Efficient Modulation", IEEE Communications Magazine, July 1979.
- [5] Gronemeyer, S. and McBride, A., "MSK and Offset QPSK Modulation", IEEE Transactions on Communications, Vol.Com-24, No.8, August 1976.
- [6] Charles Stark Draper Laboratories, Inc. Report No. CSDL-R-2167, Robust Digital Adaptive Transceiver (RDAT), RADC Contract No. F30602-87-C-1049, July 31, 1989.
- [7] Charles Stark Draper Laboratories, Inc. Report No. CSDL-R-1787, Digital Density Detector (D^3) Development Program, RADC Contract No. F30602-83-K-0160, July 31, 1985.

Mini Grant Follow-on
to the
1988 USAF-UES Summer Faculty Research Program

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by

Universal Energy Systems, Inc.

FINAL REPORT

DEVELOPMENT OF A SYSTEM TO DEPOSIT THIN FILMS OF TITANIUM CARBIDE
USING ATOMIC LAYER EPITAXY

Prepared by: Kenneth L. Walter, Ph. D.
Academic Rank: Associate Professor
Department and Chemical Engineering Department
University: Prairie View A&M University
Location: Prairie View A&M University
Date: 27 June 1991
Contract No.: F49620-88-C-0053/SB5881-0378

DEVELOPMENT OF A SYSTEM TO DEPOSIT THIN FILMS OF TITANIUM CARBIDE
USING ATOMIC LAYER EPITAXY

by

Kenneth L. Walter

ABSTRACT

Atomic layer epitaxy is a technique applicable in certain chemical vapor depositions which deposits one atomic atom layer at a time upon the substrate. This technique has been used to deposit thin films of gallium arsenide, but has not been reported in the literature to our knowledge to have been used to deposit any titanium compounds. Because of the crystal structure of titanium carbide, it may be possible that atomic layer epitaxy might be achieved under favorable conditions. If successful, precisely controlled film thicknesses could be produced with electronic, mechanical, and corrosion resistant applications.

A system has been designed and equipment ordered to construct a thin film deposition system to test this hypothesis. Funds from a current NASA grant have also been used in addition to the funds for this project. Two undergraduate senior project papers resulted from this effort. Work to assemble and test the system is still ongoing at the present, and so no concrete results are available now. Whatever results are obtained will be reported to the Air Force as soon as they are analyzed.

ACKNOWLEDGMENTS

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research, Bolling AFB, DC, as well as Rome Air Development Center for their sponsorship of this research. I also wish to thank Universal Energy Systems, Rodney Darrah and Susan Espy and their staff for their excellent job of administering this program.

I am also grateful to Dr. David Weyburne, my technical focal point at Hanscom AFB, for his key assistance in helping to initiate this project.

I. INTRODUCTION:

During a ten-week period in the summer of 1988, I participated in the 1988 USAF-UES SUMMER FACULTY RESEARCH PROGRAM/GRADUATE STUDENT RESEARCH PROGRAM. I was selected to work at the Rome Air Development Center Solid State Sciences Directorate of the Electromagnetic Materials Technology Division at Hanscom AFB, Bedford, MA, near Boston. My assignment, as jointly agreed upon with my technical focal point there, was to assist in the design, fabrication, construction, and assembly of a chemical vapor deposition system to deposit titanium nitride thin films. The particular emphasis was to attempt to achieve films deposited one atom layer at a time. This technique is known as atomic layer epitaxy. If successful, atomic layer epitaxy can be used to deposit films of exactly controlled thickness having applications in microelectronics, electrooptics, high strength materials, and corrosion resistance. The construction of this system was not complete when my research period was over, and was left to base personnel to complete and operate.

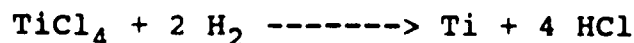
Later, I was successful in my application for a mini-grant follow-on project to develop a similar atomic layer epitaxy system on the campus of Prairie View A&M University. The focus of this application was titanium carbide, but other compounds of titanium and other elements were also possible candidates.

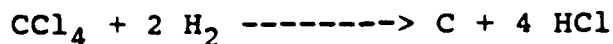
This report focuses on the efforts to design and construct such a system. Funds from NASA, obtained in a separate proposal, were also used in developing the system. That NASA grant still continues.

II. OBJECTIVES OF THE RESEARCH EFFORT:

Gallium arsenide (GaAs) has frequently been deposited using atomic layer epitaxy, as have other II-VI and III-V semiconductor compounds. GaAs has the zincblende crystal structure, as shown in Figure 1, and it is easy to visualize how a layer of Ga could be followed by a layer of As, and so forth, until a thin film of several atom layers of each element was deposited to make up the compound. The structure of titanium carbide, where the small carbon atom is at the center of an octahedral structure, is shown in Figure 2. Although it is difficult to imagine how a single layer of titanium atoms could be followed by a single layer of carbon atoms to make up the structure, it is precisely this that we are attempting to accomplish.

In short, it is our objective to design, construct, and assemble the components necessary to attempt the atomic layer epitaxial deposition of titanium carbide (TiC). We anticipate that this can best be done at a substrate temperature in the vicinity of 750°C and at a pressure of approximately 40 torr. The source of the titanium is to be titanium tetrachloride (TiCl₄), which is a liquid at room temperature. The titanium chloride will be vaporized by and carried to the reactor by a stream of hydrogen. Vaporization will occur in a glass bubbler. The carbon source is to be carbon tetrachloride (CCl₄) liquid, also vaporized by hydrogen in a bubbler. Hydrogen serves as the carrier gas, and also serves as a reactant in the reactions





Then the titanium carbide is formed on the substrate surface by



The key to having these reactions proceed one atom layer at a time rests with the reaction mechanism. If the ease of vaporization of a complex carrying a titanium atom is greater than the ease of forming a titanium metal layer on top of TiC, and the same is similarly true for a carbon-containing complex, then there is a good chance that atomic layer epitaxy (ALE) will occur. If ALE does occur, then precise control of layer thickness can be obtained, in fact precise control to the nearest single atom layer.

The procedure will be as follows:

- 1) Evacuate the chamber containing the surface-cleaned substrate, and heat the substrate to the desired deposition temperature.
- 2) Flush the system with pure hydrogen, and adjust the system pressure to the desired value.
- 3) Allow the gas flow containing Ti to enter the chamber, and contact with the substrate occurs for a few measured seconds.
- 4) Flush the chamber with hydrogen to remove nearly all of the Ti-containing gas, and desorb any unreacted Ti-complex from the substrate. This may take many multiples of the time period of step 3.
- 5) Admit the C-containing gas for a measured time interval.
- 6) Repeat step 4 for the C-containing gas.
- 7) Return to step 1 until the desired number of monolayers are

deposited.

III. LITERATURE REVIEW:

The literature review of the proposal is included as an appendix, since it would serve no purpose to repeat it here. Most of it serves to chronicle the chemical vapor deposition of electrically conductive titanium compounds -- titanium boride, titanium carbide, and titanium nitride. None of these compounds had been reported in the literature as deposited by atomic layer epitaxy as of January, 1989, and the same still holds true today.

What will be discussed here is the additional literature since the original proposal. All of these relate to titanium nitride, except for the senior papers of the two undergraduate students who have worked on the project reported upon here.

Ianno, et al. (1) used plasma enhancement to deposit titanium nitride using nitrogen as the nitrogen source. They used a rather high residual pressure of 1.1 torr. They were able to achieve reasonably low resistivity thin films at substrate temperatures as low as 500°C. They also investigated changing flow rates of the reactants, and found that nitrogen had to be supplied in excess of stoichiometric amounts to reduce the chlorine content of the films. The chlorine content is mainly responsible for any increase in resistivity. They proposed a mechanism where $TiCl_3$ was the reactive species. Several uses of these films as metal-lization barriers in microelectronic applications were given.

Ishihara et al. (2) produce TiN films using a metalorganic source for the titanium -- $\text{Ti}(\text{N}(\text{CH}_3)_2)_4$. In this

case, ammonia was used as the nitrogen source, and the pressure was 0.3 torr. The temperature range investigated was 300-580°C, but the golden TiN color did not develop until 580°C, where the lowest resistivity films were obtained. Step coverage was good, approximately 50%.

Yokoyama, et al. (3) use the same chemical system as Ishihara, et al., but at a pressure of only 0.188 torr. They developed a new heater system capable of 800°C, but carried out depositions in the range 500-700°C. In this case, step coverage was almost perfect, even filling completely some large aspect ratio, small diameter holes. They also carried out hydrogen annealing of the deposited films. This technique dropped chlorine content from the small deposited level of 5.8% to below 1%, and resistivities also dropped. The authors suspect that the chlorine is removed as ammonium chloride, which explains why excess nitrogen is beneficial. Their 500°C deposit is dark red-brown, but changes to the typical golden color upon annealing. The deposit is highly columnar.

Michelle Boyard (4) and Caryn Davis (5) each wrote Senior Design Project Reports based on the work they accomplished in designing portions of the hardware for our deposition system. Miss Boyard's report concentrates on the temperature control and the plasma control systems, including the heaters and the plasma system's matching network. Plasma may be used in this work, and

will definitely be used for NASA's amorphous silicon deposition work. Miss Davis' report discusses the design of the gas supply system and the vacuum system.

IV. DESIGN OF THE HARDWARE:

The main equipment item is the deposition chamber, depicted schematically in Figure 3. It consists of a vycor glass cylinder 13 inches in diameter and about 16 inches high. The chamber fits over a stainless steel support ring, through which electrical and gas feedthroughs pass. Vacuum is applied from below. The substrate is mounted upon a stainless steel plate insulated from below by a ceramic, and heated by a resistance plate heater. The stainless steel can serve as an electrode should plasma depositions be required, and a heated top electrode is also present.

Figure 4 shows the schematic diagram of the gas supply system. Both feed gases are supplied in a hydrogen mixture which passes through bubblers. The reactive gases are set up in a "vent-and-run" fashion, which means that the flow rate through the bubblers is maintained constant. Each needle valve is adjusted so that the path of the gas straight to the vacuum system has exactly the same flow resistance as the path of the gas through the ALE chamber. Nitrogen is available as a purge gas, since it is anticipated that ammonia would be used as a nitrogen source should TiN ever be deposited.

Figure 5 shows the overall arrangement of equipment. Mass

flow controllers maintain critical gas flow rates constant. A Tylan flow sequencer can be programmed to maintain the complicated gas flow sequence schedule required by the atomic layer epitaxy procedure.

Table 1 shows most of the equipment specifications. A silane gas supply system will also be added to the system shown, since NASA is sponsoring a concurrent project to deposit amorphous silicon thin films. Table 2 shows costs of some of the Table 1 equipment, as well as model numbers and suppliers. No particular equipment recommendations can be made, since the system has not yet been completely assembled or operated.

V. RESULTS:

As mentioned before, the system is still under construction and assembly, so no results or recommendations can be made at this time.

REFERENCES

(Note: For references from the 1989 proposal, see the Appendix.)

1. Ianno, N. J., A. U. Ahmed and D. E. Englebert, "Plasma-Enhanced Chemical Vapor Deposition of TiN from $\text{TiCl}_4/\text{N}_2/\text{H}_2$ Gas Mixtures," *Journal of the Electrochemical Society* 136(1):276-280, 1989.
2. Ishihara, Kazuya, Katsumi Yamazaki, Hidenao Hamada, Koichi Kamisako and Yasuo Tarui, "Characterization of CVD-TiN Films Prepared with Metalorganic Source," *Japanese Journal of Applied Physics* 29(10):2103-2105, 1990.
3. Yokoyama, N., K. Hinode and Y. Homma, "LPCVD Titanium Nitride for ULSIs," *Journal of the Electrochemical Society* 138(1):190-195, 1991.
4. Boyard, Michelle M., "Development of a System to Deposit Thin Films of Amorphous Silicon by Plasma-Enhanced Chemical Vapor Deposition," 1989. An unpublished senior project paper available in the Chemical Engineering Department of Prairie View A&M University, Prairie View, Texas.
5. Davis, Caryn, "Development of a System to Deposit Thin Films of Titanium Carbide by Utilizing Atomic Layer Epitaxy," 1989. An unpublished senior project paper available in the Chemical Engineering Department of Prairie View A&M University, Prairie View, Texas.

LITERATURE REVIEW WHICH ACCOMPANIED 1989 PROPOSAL

Holleck (5,6) offers review papers on selecting materials useful for hard, wear resistant coatings. He emphasizes that substrate and coating must have similar thermal expansion coefficients, and lists titanium nitrides, borides and carbides as candidate materials. Knotek (10) et al. investigate sputtered films of titanium nitrides, carbides and mixtures for hardness. Tsakalos (24) shows that superlattices with wavelengths of the order of about one nanometer show improved elastic moduli, which show maxima at certain wavelengths. Takahashi and Kamiya (21) used CVD to deposit Ti-V-B films.

Helmersson et al. (4) show a hardness maximum in the TiN/VN strained-layer superlattice system at a

wavelength of about 6 nm. They also show a maximum in hardness when the ratio of TiN layer thickness to total layer thickness is about 0.35. They used a sputtering technique to grow the films at about 750°C.

Kim and Chun (9) used CVD to deposit TiN onto TiC at about 900-1150°C and Itoh (7) deposited alternating TiN and TiB₂ layers at the same temperatures. High-hardness, adherent coatings were achieved. Moore et al. (13) produced four sequential layers on a WC cutting tool using CVD, including TiN and TiC. Kim et al. (8) discuss CVD of TiC alone.

Kurtz and Gordon (11) discuss the use of TiN in electronic and several other applications, and deposit films from 400-700°C, with 600°C being the lower limit for best results. They also test metal-organic gas sources.

Titanium boride films are complicated by the fact that the boron component can be lost by reaction with residual water vapor and subsequent volatilization, as shown by Feldman et al. (3). The result is often uncertain stoichiometry. Lynch et al. (12) showed this was a problem even in bulk single crystals.

Work on the TiB₂ system dates back to least as far as 1949, when Ehrlich (2) prepared several titanium

compounds by sintering. Shappiro et al. discuss some of the electronic applications for TiB_2 . Caputo et al. (1) and Motojima et al. (14) discuss erosion resistance and hardness of CVD coatings of TiB_2 . Pierson et al. (17,18) used CVD to produce TiB_2 coatings on graphite from 600-925°C. Takahashi and Itoh (20) used an ultrasonic field during CVD to decrease the grain size of the deposit. Williams (27) uses a plasma to lower the CVD deposition temperature of TiB_2 films.

Ozeki et al. (16) and Tischler and Bedair (22,23) give explanations of why monolayers can be deposited by ALE in the GaAs system and the InAs and In-Ga-As systems. Nelson (15) describes ALE applied to ZnS and ZnSe.

Wang et al. (25) show how TiO_2 smoke particles allow visualization of flow patterns, and assist in the spatial design of CVD reactors.

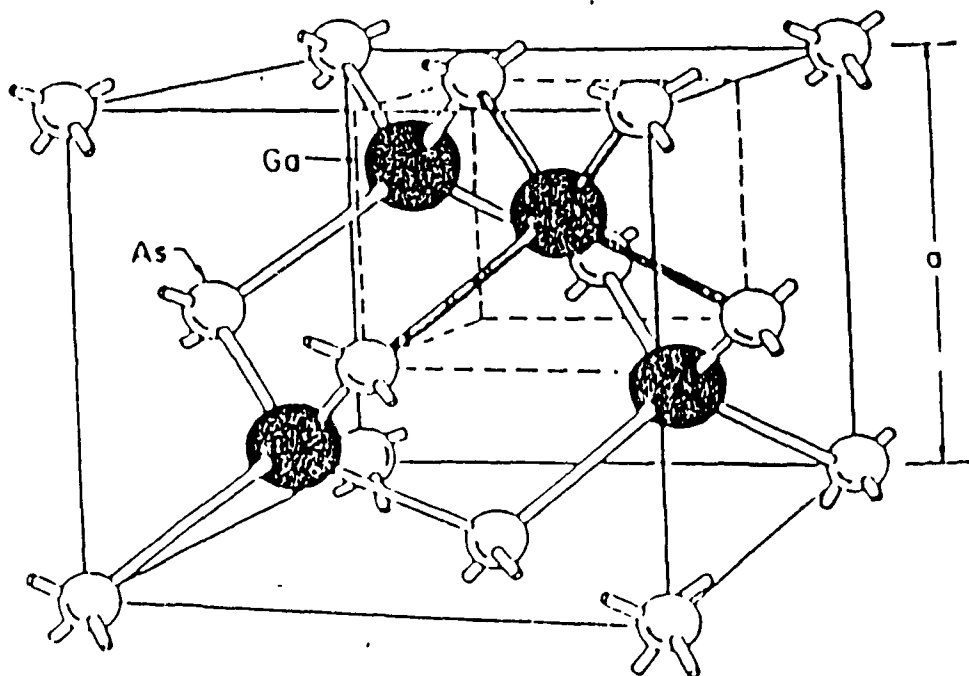
D. References

1. Caputo, A.J., W.J. Lackey, I.G. Wright and P. Angelini, "Chemical Vapor Deposition of Erosion-Resistant TiB_2 Coatings," J. Electrochemical Society: Solid State Science and Technology 132(9), pp. 2274-80, 1985.
2. Ehrlich, Paul, "Über die binären Systeme des Titans mit den Elementen Stickstoff, Kohlenstoff, Bor und Beryllium," Zeit. für Anorg. Chemie 259, pp. 1-41, 1949.

3. Feldman, Charles, Frank G. Satkiewicz and Gerard Jones, "Preparation and Electrical Properties of Stoichiometric TiB₂ Thin Films," J. Less Common Metals 79, pp. 221-235, 1981.
4. Helmersson, U., S. Todorova, S.A. Barnett and J.-E. Sundgren, "Growth of single-crystal TiN/VN strained-layer superlattices with extremely high mechanical hardness," J. Appl. Phys. 62, pp. 481-484, 1987.
5. Holleck, H., "Material selection for hard coatings", Vac. Sci. Technol. A4, pp. 2661-2669, 1986.
6. Holleck, H., Ch. Kuhl, and H. Schulz, "Summary Abstract: Wear resistant carbide-boride composite coatings," J. Vac. Sci. Technol. A3, pp. 2345-2347, 1985.
7. Itoh, Hideaki, "Effect of TiB₂ Interlayer on the CVD of an Amorphous Titanium Nitride Film," J. Crystal Growth 57, pp. 456-458, 1982.
8. Kim, D.G., J.S. Yoo and J.S. Chun, "Effect of deposition variables on the chemical vapor deposition of TiC using propane," J. Vac. Sci. Technol. A4, pp. 219-221, 1986.
9. Kim, Moo Sung and John S. Chun, "Effects of the Experimental Conditions of Chemical Vapor Deposition of a TiC/TiN Double Layer Coating," Thin Solid Films 107, pp. 129-139, 1983.
10. Knotek, O., M. Bohmer and T. Leyendecker, "On structure and properties of sputtered Ti and Al based hard compound films," J. Vac. Sci. Technol. A4, pp. 2695-2700, 1986.
11. Kurtz, S.R., and R.G. Gordon, "Chemical Vapor Deposition of Titanium Nitride at Low Temperatures," Thin Solid Films 140, pp. 277-291, 1986.
12. Lynch, C.T., S.A. Mersol and F.W. Vahldiek, "The Microstructure of Single-Crystal Titanium Diboride," J. Less Common Metals 10, pp. 206-219, 1966.
13. Moore, R.L. and L. Salvati, Jr., "Surface analysis of diffusion zones in multiple chemical vapor deposition coatings," J. Vac. Sci. Technol. A3, pp. 2425-2431, 1985.
14. Motojima, Seiji, Masahiko Yamada and Kohzo Sugiyama,

- "Low-Temperature Deposition of TiB₂ on Copper and Some Properties Data," J. Nuclear Materials 105, pp. 335-337, 1982.
15. Nelson, Jeffrey G., "Epitaxial growth of ZnS and ZnSe on the low index faces of GaAs using atomic layer epitaxy," J. Vac. Sci. Technol. A5(4), pp. 2140-2141, 1987.
 16. Ozeki, M., Mochizuki, N., Ohtsuka, N., Kodam, K., "Kinetic processes in atomic-layer epitaxy of GaAs and AlAs using a pulsed vapor-phase method," J. Vac. Sci. Technol. B, Vol. 5, No. 4, pp. 1184-1185, 1987.
 17. Pierson, H. O., and Mullendore, A.W., "The Chemical Vapor Deposition of TiB₂ From Diborane," Thin Solid Films 72, pp. 511-516, 1980.
 18. Pierson, H.O., E. Randich and D.M. Mattox, "The Chemical Vapor Deposition of TiB₂ on Graphite," J. Less Common Metals 67, pp. 381-388, 1979.
 19. Shappirio, J.R., J.J. Finnegan and R.A. Lux, "Diboride diffusion barriers in silicon and GaAs technology," J. Vac. Sci. Technol. B4, pp. 1409-1411, 1980.
 20. Takahashi, Takehiko, and Hideaki Itoh, "Ultrasonic Chemical Vapor Deposition of TiB₂ Thick Films," J. Crystal Growth 49, pp. 445-450, 1980.
 21. Takahashi, Takehiko, and Hideo Kamiya, "Chemical vapor deposition of the system Ti-Zr-B," High Temp.-High Press. 9, pp. 437-443, 1977.
 22. Tischler, M.A. and S.M. Bedair, "Self-limiting mechanism in the atomic layer epitaxy of GaAs," Appl. Phys. Lett. 48, pp. 1681-1683, 1986.
 23. Tischler, M.A. and S.M. Bedair, "Improved uniformity of epitaxial indium-based compounds by atomic layer epitaxy," Appl. Phys. Lett. 49, pp. 274-276, 1986.
 24. Tsakalakos, T., "Mechanical properties and diffusion of metallic superlattices," J. Vac. Sci. Technol. B4, pp. 1447-1456, 1986.
 25. Wang, C.A., S.H. Groves and S.C. Palmateer, "Flow Visualization Studies for Optimization of OMVPE Reactor Design," J. Crystal Growth 77, pp. 136-143, 1986.

26. Walter, Kenneth L., "Chemical Vapor Deposition of Titanium Compounds with an Atomic Layer Epitaxy System," Final Report submitted to Universal Energy Systems and the U. S. Air Force for the 1980 USAF-OBSS Summer Faculty Research Program.



Zinc blende

Figure 1. Zincblende lattice (GaAs)

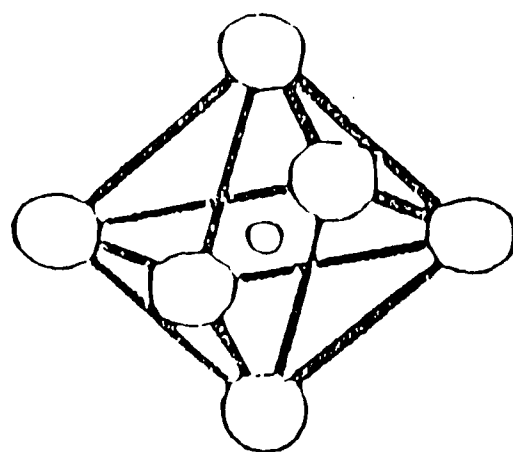


Figure 2. Crystal Structure of Titanium Carbide

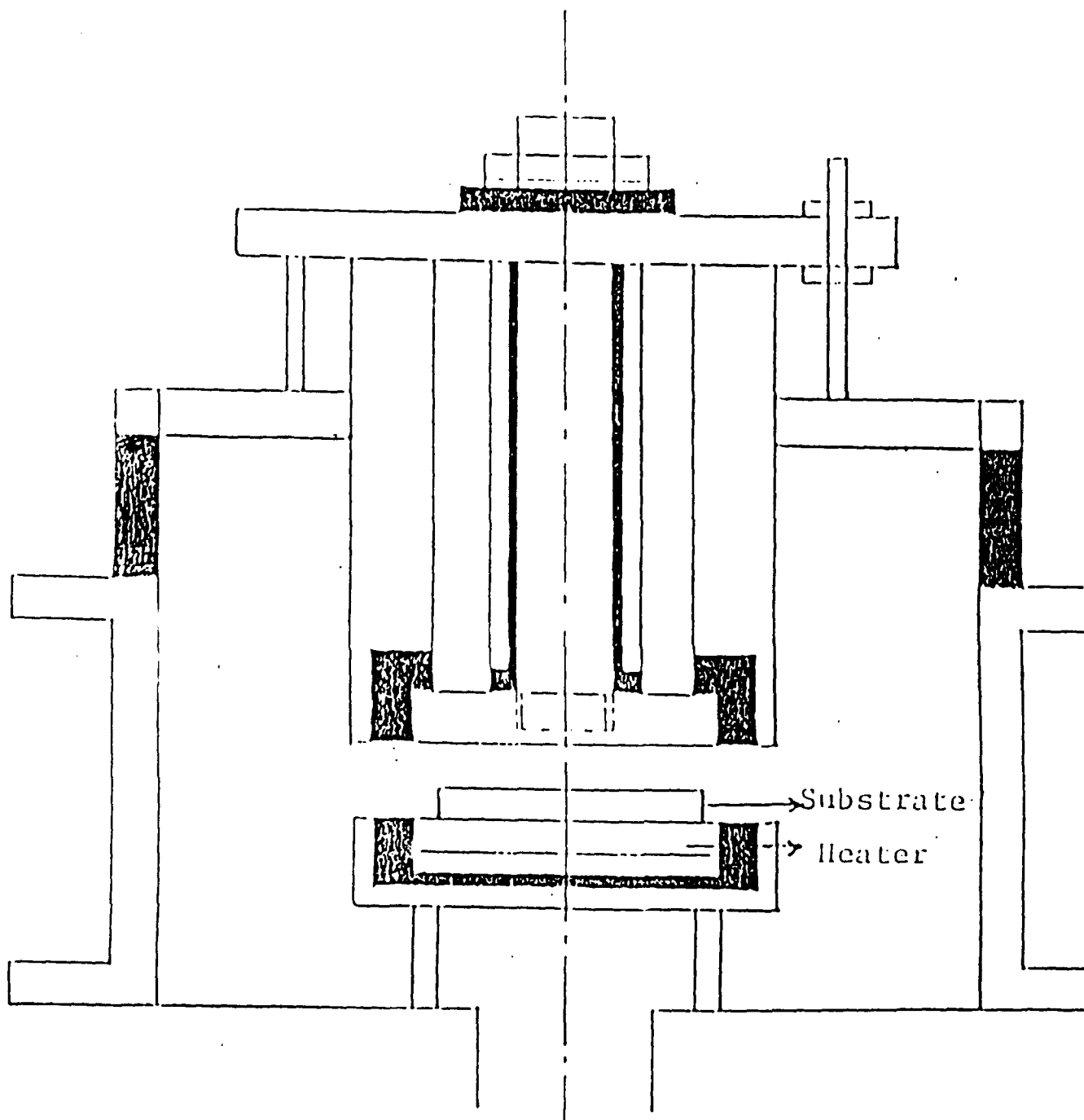


Figure 3. Atomic Layer Epitaxy Chamber

SCHEMATIC AND CONCEPTUAL DESIGN OF ALE SYSTEM

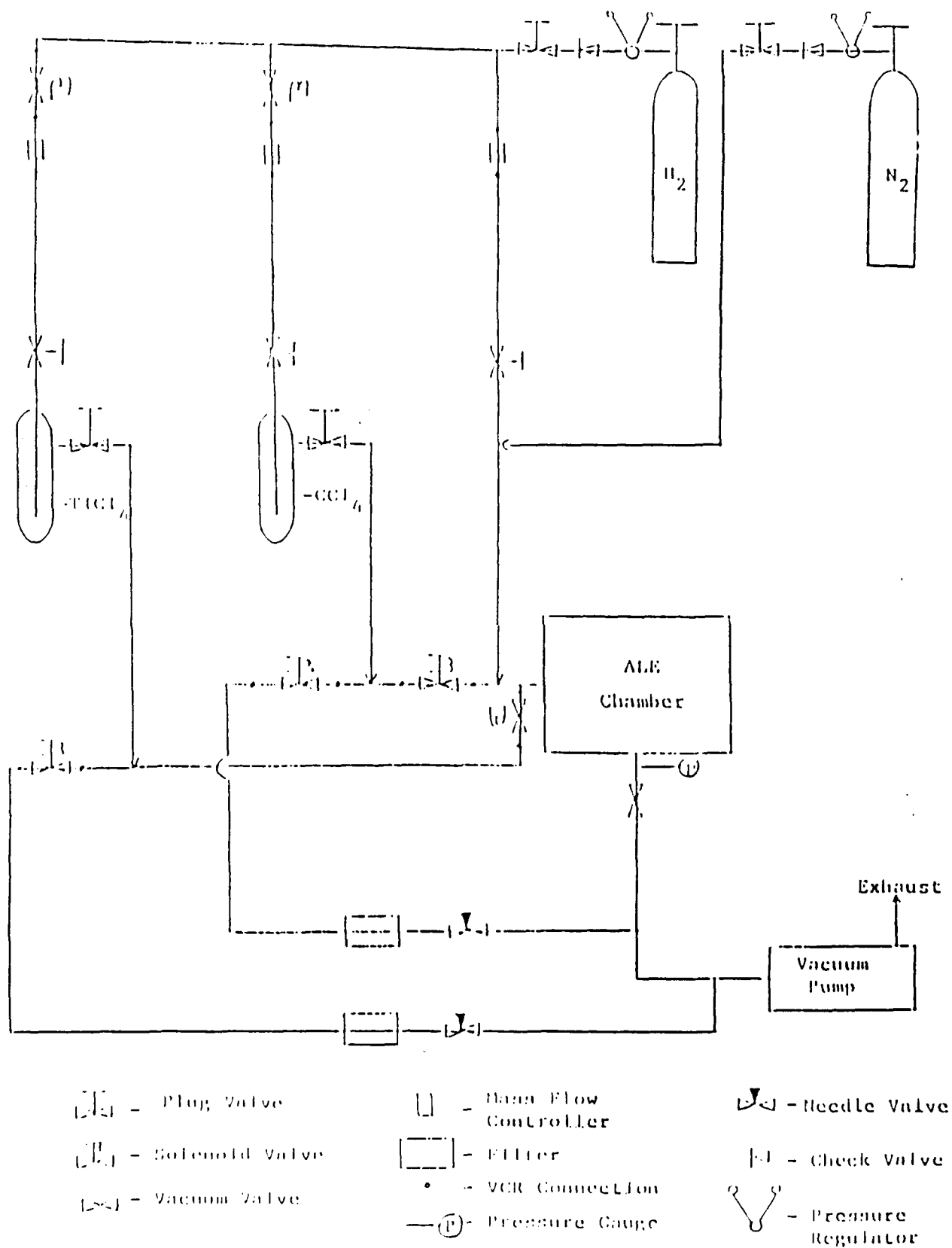


Figure 4. Schematic of ALE System

Inside of Hood

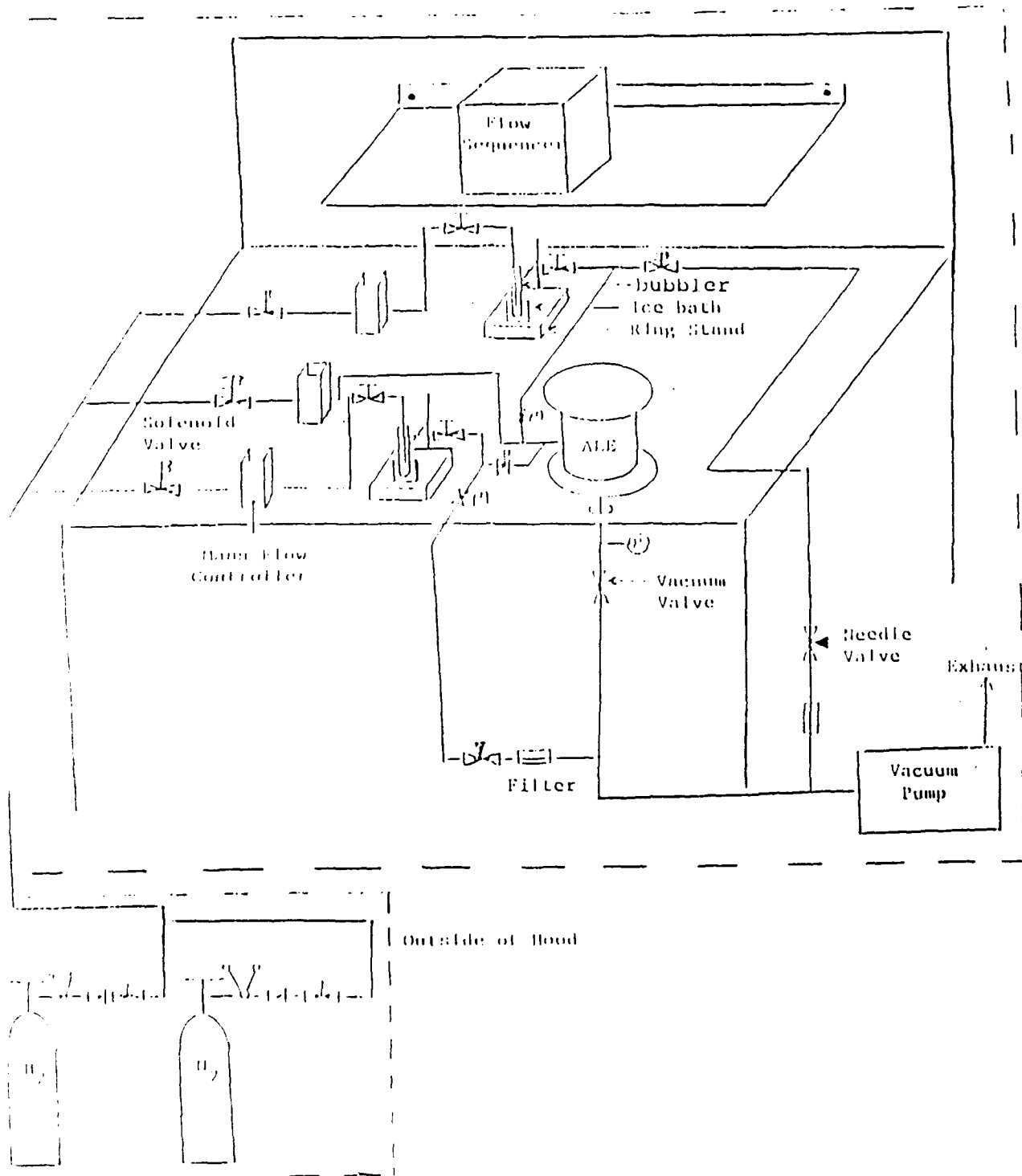


Figure 5. Conceptual Design of ALE System

TABLE 1.
DESIGN SPECIFICATIONS FOR EQUIPMENT

Equipment	Dimensions	Parameters	Materials of Construction
ALE Chamber	D ~ 13 in. H ~ 15.5 in. V ~ 33.8 L		Aluminum baseplate, machinable ceramic insulators, glass side wall
Bubblers	D ~ 1 in. L ~ 6 in. V ~ 30 ml		glass
Vacuum Pump $P_v = 1/2$ hp	L ~ 19.25 in. W ~ 14.12 in. H ~ 15.38 in.	$S_{ma} = 525$ rpm $P_{mi} = 1 \times 10^{-4}$ Torr	stainless steel
H ₂ gas cylinder	D ~ 9 in. L ~ 51 in. $P_d = 20$ psig	$P_{ma} = 2,200$ psig $m_{ma} = 545.8$ g	carbon steel
H ₂ gas cylinder	D ~ 9 in. L ~ 51 in. $P_d = 20$ psig	$P_{ma} = 2,490$ psig $m_{ma} = 8,950$ g	carbon steel
SiH ₄ gas cylinder	D ~ 2 in. H ~ 13 in.	$P_{ma} = 400$ psig $m_{ma} = 17$ g	carbon steel
H ₂ pressure regulator	$P_d = 4-100$ psig	two-stage	body-brass
H ₂ pressure regulator	$P_d = 4-100$ psig	two-stage	body-SS
SiH ₄ pressure regulator	$P_d = 0-25$ psig	single stage	body-SS
Mass flow controllers	H ~ 5 in. W ~ 1 in. L ~ 5 in. $D_t = .25$ in.	F ~ 40 sccm and 3000 sccm hydrogen	seals-VCR sapphire ball valve seat
Tymgard Flow Sequencer and Timer	H ~ 10.78 in. W ~ 6 in. Dc ~ 6.25 in.	$t_{mic} = 1$ sec. $t_{ms} = 99$ steps/ recipe	

TABLE 1 , continued

Vacuum valve	D = 1 in.		body-SS copper O-ring
Solenoid valves	D = .25 in.	normally closed	body-SS
Needle valves	D = .25 in.	O = .055 in. ₃ F _m = 2510 cm ³ /s N _{m'l} = 6	body-SS seals-SS
Plug valves	D = .25 in.		body-SS seals-SS
VCR connections	D = .25 in.		body-SS seals-SS
Elbows	D = .25 in.		all SS
Unions	D = .25 in.		all SS
Tees	D = .25 in.		all SS
In-line filter		D _p = .5 x 10 ⁻⁶ m	Body-SS filter medium - mesh element
Check valves		P _O = 20 psig	Body-SS O-ring- viton
Vacuum pump oil		V _p = 2.1 L	Fomblin
Pressure gauge	D _t = .25 in.	P = 0-760 Torr	

Table 2. Price List

Model, type or part no.	Component	Quantity	Manufacturer	Total Price
SS-404-1	Ferrules (back)	50	Nupro	204.00
SS-403-1	Ferrules (front)	50	Nupro	30.50
Tymgard	Flow sequencer and timer	1	Tylan	3430.00
1A	H ₂ gas cylinder	1	Matheson	\$100.00
3104-350	H ₂ regulator	1	Matheson	\$286.00
SS-4F-.5	In-Line Filter	2	Nupro	\$72.80
FC-260	Mass Flow controller	4	Tylan	\$3580.00
7532-10	Midget bubbler	2	Ace Glass	\$78.00
SS-4BMW	Needle Valve	2	Nupro	\$1738.00
1A	N ₂ gas cylinder	1	Matheson	\$35.00
0-500	N ₂ regulator	1	Matheson	\$196.00
S-4H	Plug Valve	20	Nupro	\$1738.00
25-1009SW-02L	Pressure Gauge	1	Myer	\$48.51
7x	SiH ₄ gas cylinder	1	Matheson.	\$147.00
3601-350	SiH ₄ regulator	1	Matheson	\$517.00
SS-400-6	Union	20	Swagelok	\$143.00
SS-400-9	Union Elbow	20	Swagelok	\$204.00
SS-400-3	Union Tee	40	Swagelok	\$602.00
Fomblin	Vacuum Pump Oil	1	Lesker	\$875.00
304-24VFBG	Vacuum Valve	1	Nupro	\$525.30

Note: Swagelok and Nupro parts were purchased through local supplier,
North Houston Valve and Fitting Company.

1990 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

conducted by

UNIVERSAL ENERGY SYSTEMS

FINAL REPORT

Neural Networks for Invariant Pattern Recognition

Prepared by:	James S. Wolper
Academic Rank:	Assistant Professor
Department:	Department of Mathematics and Computer Science
University:	Hamilton College
Date:	8 December 1990
Contract Number:	F49620-88-C-0053

I. Introduction

Pattern Recognition is an important problem for both military and civil applications. The difficult part of the problem is that real-world patterns are generally transformed in many ways before being presented for identification. The ability to recognize transformed patterns is called Invariant Pattern Recognition.

Neural Network architecture is a relatively new technique. (Actually, the technique is fairly old, but it was only recently that its value has become widely recognized.) Neural Networks exhibit good behavior with respect to incomplete data, they offer the speed advantages of parallel implementation, and they have "graceful degradation"; this means that a Neural system can continue to perform well in the event of component failures, rather than just stopping. These traits suggest that Neural Networks may be important parts of various pattern recognition systems. This was the general focus of this research effort.

The particular area of application was image recognition. The images used were gray-scale images of aircraft planforms, simulating overhead imagery. The system was 'trained' on these images, and tested on images which had been transformed by translation, scaling, occlusion, small in-plane rotations, and additive Gaussian white noise of fairly high power. The objective was to design a Neural Network based system which could successfully recognize the transformed images.

II. The Problem

An image is a finite energy (ie, L^2) function $I: \mathbb{R}^2 \rightarrow \mathbb{R}$. The value of $I(x, y)$ is the intensity of the image at the point (x, y) . It can be assumed that the value of I is 0 outside of a bounded set, ie, there are no images of infinite extent.

Research into the Human Vision System (HVS) (see [KFAW] for recent ideas) indicates that it uses several scales in the identification process (the system is, of course, much more complicated than this, although multiscale is an important strategy). At the largest scale it determines gross attributes (house, airplane, bird, Chinese character) of the objects which it senses. For example, the large scale view of an object in a overhead image might determine that it is a wide-bodied airplane; a slightly smaller scale view might determine that it has two wing-mounted engines, thus reducing the possible solutions to Airbus A320 or Boeing 767; and a smaller scale view might detect the presence of flap track fairings, which are not present on the 767. (Of course, at very small scale one might be able to read the manufacturer's

data plate.)

Kosslyn et. al. call some of these small-scale features "trigger features". The presence of a "trigger feature" suffices to identify an object despite gross transformations. A musical analogy might be the ability to name certain popular songs after hearing just a few notes.

It seems reasonable to design an automatic identification system which mimics this approach (although, as will be discussed later, this is not the only good approach). The problem is to use the various scales in a systematic and manageable way, and to determine the appropriate "trigger features".

III. Design Considerations

The original plan for this effort was based on a Neural Network architecture called the *Neocognitron*. The neocognitron is a multiple-layer feed forward network. Each layer consists of two types of units. One type of unit responds to a low-level feature in a certain region of the previous layer, while the second type 'integrates' responses from the first type to form a part of a feature at this level. This design accomplishes the multiple scale analysis described above, although there is no strict analog of the "trigger feature". Fukushima [F] developed this architecture and was successful using it to recognize handwritten numerals.

The proposal was to enhance this architecture through preprocessing of the images and other means to enable it to handle larger, gray-scale images, rather than outlines. The enhancements developed, which are described below, were such that the only vestige of the neocognitron in the system is the presence of a feed-forward network. The response to low level features has been replaced by the Gabor transform, which filters for spatial frequency and orientation; the "integration" of low level features is bypassed using entropy (low entropy indicates high information content), which selects the *characteristic features* in the image; and the slowly-trained neocognitron has been replaced by the Probability Neural Network, which features instantaneous training. The resulting system is able to work easily with gray scale images but is more flexible and easier to train.

IV. Implementation Considerations

All of the algorithms discussed in this report were implemented in C programs which simulate parallelism. Programs were primarily developed on the Macintosh

Itci computer which was purchased with the grant funding. The programs have been successfully ported to a Sun workstation and IBM-PC compatibles.

Several relatively inexpensive parallel machines have been announced since this effort began. These include the MasPar and DAP machines. It would be very interesting to test these algorithms on such a machine, especially in view of the parallel nature of the algorithms.

It had been proposed to test these algorithms with "real" images, but security and funding considerations prevented this.

V. The Design

The first step in the processing of the images is the "Gabor transform". The Gabor transform is analogous to the Fourier transform in the sense that it forms a representation of a function such as $I(x, y)$. Given vectors k and x , define

$$G(k, x) = \exp(-\|k\|^2 \|x\|^2 / 8.77) \exp(-ik \cdot x)$$

where $\|x\|$ denotes the magnitude of the vector x , $i^2 = -1$, and 8.77 is a parameter controlling window size. The first factor is a Gaussian; the second is sensitive to frequency and orientation. One may consider G to be the transfer function of a 2-D filter.

The transform consists of a convolution of I with $G(k, x)$ for various k at points on a grid superimposed on the image. This is analogous to filtering with a bank of filters. The filter responses form the *Gabor transform*.

By proper choice of the various k , we can search for various features. If $\|k\|$ is large, the filter will have a high response in the presence of correctly oriented high-frequency features; if $\|k\|$ is small, the response is high only in the presence of low-frequency features with the correct orientation. The value 8.77 sets the window size to two wavelengths. See [D] or [FT] for more on this technique.

Notice that the filtering is local with respect to the image, due to the limited Gaussian window, so a parallel machine could filter simultaneously at all points of the superimposed grid.

Define a *characteristic feature* to be the presence of two or more low-level features at the same scale. Conceptually, a characteristic feature is a corner or the presence of several edges in a neighborhood. The S-cells of Fukushima's neocognitron essentially respond to characteristic features as defined here. Characteristic

features are used to classify objects in images. They are scale and position invariant, and can be made rotation invariant through normalization.

The system needs to select "important" characteristic features from each image which it is supposed to know. This is done using uncertainty in the sense of Information Theory ([Sh]). First, the filter responses are quantized. Let p_j be the observed probability of response j in the training set. Then the uncertainty at a given point on the image is

$$H = - \sum p_j \log_2(p_j)$$

where the sum is taken over all of the Gabor filter responses at each orientation at the grid point in question. It is possible to calculate H in parallel. A low value of H at a certain location indicates that there is something "interesting" or "informative" there. The characteristic features are those with the lowest value of H ; the lowest r values are taken, where r is a user specified parameter.

Training of the system consists of performing the Gabor transforms on all of the images, selecting the characteristic features from these images, and training some kind of Content Addressable Memory (CAM) to recognize combinations of features. (There is no reason to believe that any give characteristic features will be present in only one image.) The CAM selected was the Probability Neural Network (PNN). A PNN is topologically similar to a back-propagation network, but it is much more mathematically sophisticated. There is one hidden node for each training exemplar, as well as one for each category. The weights from the input nodes to the hidden node representing the exemplar are exactly the inputs for the exemplar; thus, training is immediate, since there is no need to wait for some training process to converge. The transfer function of the exemplar nodes is not the familiar sigmoid $\frac{e^x}{1+e^x}$; rather, it is chosen to enable each output to represent the probability density for a category, evaluated at the input in question. This gives good figures-of-merit for the classifications inferred. See [S] for more details.

VI. Testing the System

The system outlined above was simulated with C programs. The first image set consisted of four 32×32 pixel hand-coded binary images representing aircraft planforms. The choices of k gave 4 orientations at each of 2 scales, and the superimposed grid was 4×4 pixels. Quantization seems to remove variations due to

uniform changes of intensity. The grid cells whose value of H was lowest generally correspond to 'important' parts of the image, eg, wingtips, fuselage-wing junctions, and engine nacelles.

In the training phase, a list is made of the characteristic features in the images. Then, the number of each type of characteristic feature in each image is counted. This count, after normalization, is the input to the PNN; the output is a score for each category.

Here is a sample training session from the SPARCstation; the user's input is in *italics*. The aircraft planforms were those of the Douglas DC-8 transport, the F-15 and F/A-18 fighters, and the P-3 Antisubmarine Warfare aircraft. These were chosen because there is some similarity between the various types, yet each is distinct.

In the program 'info', the user is given the option of examining the characteristic features in an image constructed as follows: the intensity in each grid square is inversely proportional to the uncertainty in that square, so that the most important features are the brightest.

[JSW] *gabor*

Program to compute the Gabor transform of an image.

Image size is 32 by 32.

Where is the input file? *f15.pic*

... looking for file f15.pic

Where should I write the coefficients? *f15.gab*

Maximum modulus is 384.925415

[JSW] *info*

Program to interpret gabor transforms of images.

Where is the transform file? *f15.gab*

... looking for file f15.gab

Number of quantization levels? *8*

Where should I write the entropy? *f15.ent*

Number of grid squares to display? *5*

Write transformed images to files (y/n)? *n*

Where is the original image? *f15.pic*

... looking for file *f15.pic*

/* 3 others processed the same way */

[JSW] *feature*

Feature extraction program. This program examines the *.ent files for the training set and creates a training file for pnn.c.

2 scales and 4 orientations.

Number of features from each transform? *5*

Name of the training set *planes*

Include training narrative in output (y/n)? *y*

Image to process? *dc8*

Another image (y/n)? *y*

Image to process? *f15*

Another image (y/n)? *y*

Image to process? *f18*

Another image (y/n)? *y*

Image to process? *p3*

Another image (y/n)? *n*

Training file contains 24 exemplars.

To identify an unknown image, its Gabor transform is calculated and the image's characteristic features are determined. These are compared with the known characteristic features. The (normalized) feature counts are the input to the Probability Neural Network.

With the first training set, the following unknown images were successfully identified: the left half of an F-15, the shifted top (nose) portion of an F/A-18, a shifted DC-8, various images subjected to slight in-plane rotations, and various images with additive white Gaussian noise of fairly high power (SNR approximately 10

dB). The system recognized slightly rotated images even though the normalization for rotation-invariance was not implemented.

Another training set consisted of three 32×32 pixel digitized images of human faces with 16 scales of gray. The system recognized the faces despite various occlusions and noise, although more levels of quantization were needed to be successful. The system was also run on 64 pixel images of human faces with 256 shades of gray, thus demonstrating its scalability.

[Note: the implementation of the Gabor transform and of the entropy calculations was performed during the term of this grant, as written in the original budget. Also, much experimentation was done with the Probability Neural Network algorithm. The algorithm for feature selection and the final choice of a Probability Neural Network were done under the auspices of a UES Summer Faculty Research grant at Rome Air Development Center, Griffis AFB, NY. Following the SFRP, more work was done on the theoretical extensions to this system which are described in the following sections of this report.]

VII. Another Approach

The work above has suggested a new approach to Invariant Pattern Recognition. This approach is based on sophisticated notions from the mathematical fields of Functional Analysis and Representation Theory. These ideas will be described below.

The natural setting for questions of invariance is the theory of groups. A group can be defined abstractly, but can be best thought of in this context as a set of transformations which is closed under composition. For example, the set of rotations of the ordinary (x, y) -plane about the origin forms a group: two such rotations can be composed to yield a rotation (whose value is the sum of the angles through which the component rotations move points). There is an *identity* rotation, namely, rotation through 0° . Each rotation has an *inverse*, namely, the inverse of the rotation through θ° is the rotation through $-\theta^\circ$. The set of transformations acting on the images we would like to identify — namely rotations, translations, and scalings — also forms a group.

We say a group acts on a set if it transforms elements of the set to elements of the set. If all elements of the group G which acts on a set X are allowed to act on some element x , the set obtained is called the *orbit* of x and denoted Gx .

For example, imagine that G is the group of rotations of the plane about the origin (usually denoted $O(2)$). Then the orbit of some point (x_0, y_0) is the circle of radius $r = \sqrt{x_0^2 + y_0^2}$.

Now, suppose we have a group G acting on a set X and a set of patterns $\{I_1, \dots, I_r\}$ in X . To identify an unknown pattern in a way which is invariant under the action of G , we propose to precompute all of the orbits $\{GI_j\}$. Then the task of identification of an unknown pattern U becomes the task of finding the orbit closest to U . This procedure is automatically robust with respect to noise because the noisy version of U is probably close to U .

For example, suppose that we had patterns represented as points in the (x, y) -plane, say $(x_1, y_1), \dots, (x_r, y_r)$, and we would like to recognize these patterns even though they may have been transformed by the action of the rotation group $O(2)$. In this case, we can precompute the orbits by computing their radii: $r_i = \sqrt{x_i^2 + y_i^2}$. To identify an unknown pattern (x, y) , we compute its radius $r = \sqrt{x^2 + y^2}$, and determine which r_i is closest to r . This tells us which (x_i, y_i) was rotated to obtain (x, y) . If $r = 2$ and $(x_1, y_1) = (1, 1)$ while $(x_2, y_2) = (2, -4)$, we conclude that the pattern $(-3, 4)$ is a rotated noisy version of $(2, -4)$ since 5 is closer to $\sqrt{20}$ than it is to $\sqrt{2}$. This is impressive when one considers that $(-3, 4)$ is much closer to $(1, 1)$ than it is to $(2, -4)$ (5 versus $\sqrt{89}$).

In image recognition, it is probably not the case that we want to let X be the plane and then consider the orbits of the images we plan to remember. More likely, we can achieve significant compression (and subsequent computational improvement) by considering some representation of the group in question. A representation of a group of degree n is an $n \times n$ matrix associated to each group element in such a way that the group composition becomes matrix multiplication. It is a fundamental fact that every group has many representations of various degrees. For details consult [Serre].

One way to approach the use of group representations for Invariant Pattern Recognition is through the selection of feature vectors. Suppose that each image I_j has an r -dimensional feature vector v_j associated to it, and suppose that we have a degree r representation of the relevant group. Then the orbit of I_j can be computed by performing repeated matrix multiplication of v_j by the matrices representing the group. If we have a good 'sampling theorem' (which would tell us how many samples of the image we need for accurate reconstruction), we could in

fact consider the orbits to be a finite set of points in R^r . Each such point could be assigned to a (virtual) processing element of a suitable parallel computer, ie, each processing element would remember the name of the image in whose orbit it lies. Recognition then consists of finding the feature vector for the unknown image, then spreading out along $r - 1$ -dimensional hyperspheres until an orbit is encountered.

The amazing thing about such a system is that the speed of identification *improves* as more patterns are learned. This is because the average distance from any point to one of the patterns must go down as the density of 'known' points increases.

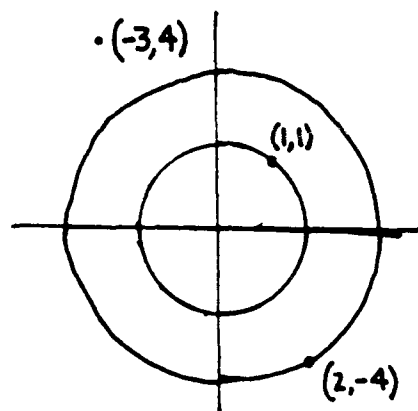


Figure 1

Implementation of such a strategy depends on an appropriate and easy-to-compute scheme for determining feature vectors. A strong candidate for such a scheme is the theory of *wavelets* (see [M]). A wavelet scheme is an orthonormal basis for $L^2(R^2)$, which means that it is easy to express a given image in terms of this basis. According to [H] there are appropriate sampling theorems, so a finite-dimensional feature vector is possible.

It may be objected that a wavelet scheme is too abstract, but in fact the wavelets have strong intuitive meanings similar to Gabor coefficients. Thus, a wavelet can be used to perform some of the low-level feature detection employed in the system discussed above as well as in the neocognitron.

In implementing such a scheme it is important to consider *irreducible* repre-

sentations. Every representation of a group G can be decomposed into a 'direct sum' of irreducible representations. Operationally this means that there is a basis for R^r such that the matrices for each group element decompose into block form. This has computational consequences: if a representation of degree $m + n$ can be decomposed into representations of size m and n , then the number of matrix entries needed drops from $m^2 + 2mn + n^2$ to $m^2 + n^2$, a saving of $2mn$.

The problem is that the some of the groups which arise in this context are not 'nice', so much theoretical work must be done to determine irreducible sub-representations. This leads to an additional problem: write computer programs to determine irreducible representations. Much combinatorial machinery from Representation Theory is available to help with this task.

Another consideration is *invariant theory*. In our example of the rotation group $O(2)$ acting on the plane, we determined that each orbit was uniquely determined by its radius. Invariant theory tries to determine such 'invariants' for other group actions. In particular, the invariant theory of the group of perspective transformations is well-understood, and there have been some efforts to apply results from this area to image recognition.

All of these ideas should be vigorously pursued on an appropriate parallel computer.

VII. Conclusions

In this grant effort, the use of Neural Network techniques in Invariant Pattern Recognition has been investigated. The neocognitron was considered, and a system based on Gabor transform preprocessing, Information Theoretic feature selection, and the Probability Neural Network was demonstrated to give very good performance. The success of this effort leads to ideas to use wavelets and representation theory to design a parallel pattern recognition system which would be very fast (with speed *increasing* when more patterns are learned), invariant, and robust with respect to noise.

References

- [D] Daugman, J., "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression", *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-36* (1988), 1169 - 1179.
- [F] Fukushima, "Neocognitron: a hierarchical network capable of visual pattern recognition", *Neural Networks 1* (1988).
- [FT] Flaton, Kenneth A. and Scott T. Toborg, "An approach to image recognition using sparse filter graphs", *International Joint Conference on Neural Networks*, 1989.
- [H] Healy, Dennis, personal communication (1990).
- [KFAW] Kosslyn, Stephen M., Rex Flynn, J. Amsterdam, and G. Wang, "Components of high-level vision: a cognitive neuroscience analysis and accounts of neurological syndromes", *Cognition 34* (1990), 203 - 277.
- [M] Mallat, S. G., "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-11* (1989), 674 - 693.
- [Serre] Serre, J-P., *Representations of Finite Groups*, Springer-Verlag Lecture Notes in Mathematics.
- [Sh] Shannon, Claude, *The Mathematical Theory of Communication*, Urbana: Illini Books (1969).
- [S] Specht, D., "Probabilistic Neural Networks", *Neural Networks 3*(1990), 109 - 118.

The Performance of IR Detectors Illuminated by Monochromatic Radiation

Brian P. Beecken
Bethel College
St. Paul, Minnesota

November 21, 1990

Abstract

Infrared detector performance can be changed when exposure to normal blackbody radiation is replaced by monochromatic radiation. The possible difference in detectivity occurs because the photon noise limit may be either raised or lowered. If the detector is operating near the BLIP limit, then its performance may be degraded if that limit is lowered significantly. Alternatively, if the BLIP limit is raised, then no significant effect on the detector's performance is anticipated. The radiation wavelength, detector cut-off wavelength, detector quantum efficiency, system bandwidth, and incident background photon flux from the lab combine to determine the magnitude and direction of the change. Such effects are in theory easily quantifiable and probably will be significant for the development of a simulated mission testing capability which uses laser radiation.

1 Introduction

Arnold Engineering Development Center in Tullahoma, Tennessee performs tests on infrared detectors that are intended for use in a wide range of missions. Currently, scientists at Arnold Center are developing a revolutionary system for testing IR detectors under various simulated mission environments. Their direct write scene generation (DWSG) system uses a laser beam to "paint" the anticipated scene onto the pixel elements of the detector array.

When doing testing under simulated conditions, the validity of the simulation must always be carefully considered. In the case of the DWSG, one area of concern is the ramifications of using a monochromatic laser to portray the multiple wavelengths of radiation inherent in virtually any realistic scene. An effort has been made to address this issue from a theoretical standpoint.

The theory, as discussed in the ensuing sections, indicates that if the detector is intended to work at or near the fundamental noise limit which arises from the quantum nature of light, then the simulation could easily be affected. It is possible to calculate beforehand when these effects are likely to occur and to what degree they will alter the simulation. The second half of this paper is devoted to such calculations.

Before the implications for the DWSG will have full meaning, their source must be thoroughly understood. The potential problems are deeply rooted in the basic theory of the ultimate limits on photon detection. Consequently, the first half of this paper is devoted to a careful exposition of the relevant theory. Although this material is not original, it should be a useful improvement over what is found in the literature. Meticulous care has been taken to use notation that will avoid the ambiguity that often plagues this subject. Also, an error common to at least two standard references has been identified. And finally, the present analysis should be complete enough to be easily understood, yet it emphasizes those facets that are of particular importance for the DWSG system.

2 Basic Theory

This section consists mainly of material derived from standard references[1-6]. It will provide the necessary foundation for the calculations which have

been performed for the DWSG system and are discussed in Part 3.

2.1 Photon Limited Detection

In this paper, we will only be considering ideal detectors. This restriction allows us to focus our attention on effects which originate in the illuminating radiation. The ideal detector has no internal noise: no amplifier noise, no thermal noise, and no noise due to defects inherent in the detector material itself. Only the fundamental noise which is caused by the discrete nature of photons remains. Because the individual photons will arrive at random times, the signal generated by the detector will likewise fluctuate randomly. If this is indeed the only noise present, then the detection is said to be *photon limited*.

Before continuing, let us define some symbols. The incident photon flux is represented by $\phi^q(\lambda)$ which has the units of photons per second. The symbol q indicates the flux is in terms of quanta of light per second rather than power. The functional dependence on λ will always be used as an explicit declaration of monochromatic radiation. In this case, all photons incident on the detector have the same wavelength λ .

The number of photoelectrons generated per second in the detector will be given by $\eta\phi^q(\lambda)$, where η is the quantum efficiency¹ of the detector. Thus, if e is the electronic charge, the average total current generated is:

$$\bar{i} = e\eta\phi^q(\lambda). \quad (1)$$

The infrared radiation incident on the detector consists of discrete quanta. Therefore, these photons will arrive at some average rate with a Poisson distribution about the average. It can be shown (Ref. 1, p. 132 and Ref. 3, p. 13) that the rms of the deviation Δi is given by the formula:

$$\sqrt{\Delta i^2} = \sqrt{2e\bar{i}\Delta f}. \quad (2)$$

This formula represents what is often called *shot noise* and was originally derived by Schottky² to describe the noise inherent in the current passing

¹We will make the common assumption that η is independent of λ below the cutoff wavelength of the detector.

²W. Schottky, *Ann. Phys. (Leipzig)* **57**, 541 (1918).

through a temperature-limited vacuum diode. We have chosen Δf to represent the electrical equivalent bandwidth of the system. The shot noise formula gives us the noise generated by the discrete nature of the photons striking the detector. This fundamental noise will always be present, even for an ideal detector on which only signal photons are incident.

2.2 Signal Photon Limited Detection

Consider a perfect detector when there is no background photon flux. The signal current i_s is given by Eq. (1). We will add the subscript s to $\phi^q(\lambda)$ in order to indicate that it is the signal photon flux $\phi_s^q(\lambda)$. Because there is no other noise source, all noise is due to the random arrival rate of signal photons. The total rms noise current i_n is given by Eq. (2) when i is i_s . Dividing the two equations yields the signal-to-noise ratio:

$$S/N = \frac{i_s}{i_n} = \sqrt{\frac{\eta \phi_s^q(\lambda)}{2 \Delta f}}. \quad (3)$$

Because the spectral noise equivalent power $NEP(\lambda)$ is the power flux $\phi_s^E(\lambda)$ when $S/N = 1$, we need to set the above equation equal to unity and solve for the photon flux:³

$$\phi_s^q(\lambda) = \frac{2 \Delta f}{\eta}. \quad (4)$$

Then we convert to power flux by using the energy hc/λ of a single photon:

$$NEP(\lambda) = \phi_s^E(\lambda) = \frac{hc}{\lambda} \phi_s^q(\lambda) = \frac{2hc \Delta f}{\eta \lambda}. \quad (5)$$

This equation for $NEP(\lambda)$ represents the best value possible for a detector. Of course in practice, a situation where no background photons exist is seldom achieved.⁴

³The term "spectral" and the dependence on λ is used to indicate explicitly that the signal providing the NEP is monochromatic.

⁴In fact, the observant reader will recognize that the linear dependence of $NEP(\lambda)$ on Δf given by Eq. (5) is not the usual relationship that is assumed for the calculation of $D^*(\lambda)$.

2.3 Background Photon Limited Detection or BLIP

If, in addition to signal photons, background photons are allowed to fall on a detector, then more photoelectrons are generated than before. These photoelectrons contain no more signal information so they only contribute to the noise:

$$\bar{i}_b = e\eta\phi_b^q(T). \quad (6)$$

But, if this background generated "noise current" \bar{i}_b were truly a DC current, it would not be a noise at all because it would simply be a constant offset to the signal current. Unfortunately, there is a Poisson distribution around the average value \bar{i}_b exactly like the fluctuation around \bar{i}_s . Thus,

$$\sqrt{\Delta i_b^2} = \sqrt{2e\bar{i}_b \Delta f} \quad (7)$$

and substituting Eq. (6),

$$\sqrt{\Delta i_b^2} = e\sqrt{2\Delta f\eta\phi_b^q(T)}. \quad (8)$$

Note that in the above equations we have indicated the background flux depends on T . Clearly, photons generated from background sources are *not* monochromatic but have a broad range of wavelengths that are given by the Planck distribution for blackbodies. If the entire background was of one temperature, then it would be T . Otherwise, the situation is more complicated and the argument T should merely be taken to indicate nonmonochromatic flux. Note also the implicit assumption that $\phi_b^q(T)$ includes photons only up to the detector's cutoff wavelength.

When the background flux generates noise which dominates the system, then the detector is said to be operating under BLIP conditions.⁵ Occasionally the term BLIP is used in a way that implies a more general meaning, such as "the best performance possible." Technically, this is incorrect. Under rare conditions, a detector can perform better than BLIP if the noise due to the signal photons is greater than the noise due to the background. Since we are currently dealing with BLIP conditions,

$$\sqrt{\Delta i_b^2} \gg \sqrt{\Delta i_s^2}, \quad (9)$$

⁵Although the exact identification of the acronym varies from one author to another, perhaps the most common is Background Limited Infrared Performance.

and

$$\phi_b^q(T) \gg \phi_s^q(\lambda). \quad (10)$$

Under BLIP conditions, because of Eq. (9), the equation for total noise current does not need to include the noise associated with fluctuations in signal current. Thus, Eq. (8) gives the total noise current:

$$i_n = e\sqrt{2\Delta f\eta\phi_b^q(T)}. \quad (11)$$

Now the signal-to-noise ratio becomes

$$S/N = \frac{i_s}{i_n} = \frac{\eta\phi_s^q(\lambda)}{\sqrt{2\Delta f\eta\phi_b^q(T)}}. \quad (12)$$

Setting $S/N = 1$ and solving for the required signal flux yields

$$\phi_s^q(\lambda) = \sqrt{\frac{2\Delta f\phi_b^q(T)}{\eta}}. \quad (13)$$

Converting from photon flux to power flux enables us to find $NEP(\lambda)$:

$$NEP(\lambda) = \frac{hc}{\lambda} \sqrt{\frac{2\Delta f\phi_b^q(T)}{\eta}}. \quad (14)$$

This expression for $NEP(\lambda)$ under BLIP conditions agrees with that obtained by Dereniak and Crowe (Ref. 2, p. 47) and Kruse et al. (Ref. 5, p. 358).⁶ However, Kingston (Ref. 3, p. 16) and Boyd (Ref. 1, p. 135) find a different dependence on wavelength. Either Kingston and Boyd are misleading in their notation, or they have erred by blurring the distinction between monochromatic photon flux and the blackbody spectrum typically exhibited by the background. Let us be perfectly clear: λ in Eq. (14) is the wavelength of the monochromatic signal photons and has no relation to the background photon flux $\phi_b^q(T)$. This point will prove to be very important later on when we do calculations for the DWSG.

The spectral detectivity $D^*(\lambda)$ is simply the reciprocal of $NEP(\lambda)$ multiplied by the square root of the detector area A and the square root of the

⁶The expression in Ref. 5 utilizes Bose-Einstein statistics. This correction is usually ignored at wavelengths shorter than 20 μm .

bandwidth. This has the effect of making the figure of merit independent of both parameters. Thus

$$D^*(\lambda) = \frac{\sqrt{A\Delta f}}{NEP(\lambda)}. \quad (15)$$

Finally we can write the expression for spectral detectivity under BLIP conditions:

$$D^*(\lambda) = \frac{\lambda}{hc} \sqrt{\frac{\eta A}{2\phi_b^q(T)}}. \quad (16)$$

Here the incident photon flux per unit area (or "flux density") is simply $\phi_b^q(T)/A$.

The above equation forms the theoretical basis for the graph of $D^*(\lambda)$ versus λ that has been immortalized by *The Infrared Wall Chart* produced by SBRC and reproduced in Fig. 1. When the cutoff wavelength of the detector λ_c is used in place of λ , then Eq. (16) delineates the line that represents the best performance possible for an ideal detector⁷ under BLIP conditions. According to the first term, the spectral detectivity increases in proportion to λ_c , however, because of the second term, it has a more complicated dependence on λ_c that comes through the background photon flux. Recall that the total background flux $\phi_b^q(T)$ is integrated up to λ_c . If the background is 300 K, then the number of background photons received increases dramatically with cutoff wavelength below 10 μm . Thus, the detectivity decreases. But, because of the shape of the Planck distribution, very few additional background photons are acquired as the cutoff wavelength is increased beyond 10 μm , and so the detectivity increases.

If, on the other hand, the cutoff wavelength of the detector is kept constant, then $\phi_b^q(T)$ is also constant. In this case, λ represents the wavelength of the monochromatic signal, and according to Eq. (16), $D^*(\lambda)$ is directly proportional to λ . Once the signal wavelength is greater than the detector's cutoff wavelength, then obviously the detectivity plummets because the signal cannot be seen. This accounts for the general sawtooth shape of the

⁷Only photovoltaic detectors can in theory achieve the performance indicated by Eq. (16). For photoconductors, $D^*(\lambda)$ must be less by a factor of $1/\sqrt{2}$, as indicated in Fig. 1. This is because photoconductors suffer from *generation-recombination noise*. Not only is there a shot noise associated with the generation of photoelectrons, but there is also a shot-like noise caused by the recombination of the photoelectrons. These two independent noise processes in effect cause the 2 in Eq. (7) to become a 4 (cf. Ref. 1, pp. 165-169 or Ref. 6, p. 62).

experimental curves for the various detectors that appear on *The Infrared Wall Chart*.

For the sake of completeness, it should be pointed out that the spectral detectivity $D^*(\lambda)$, which is obtained for a monochromatic signal source, may be converted to the blackbody detectivity⁸ $D^*(T)$. The latter is observed when the signal is provided by a blackbody source of temperature T . The appropriate conversion equations and graphs may be found in the literature[2, 4, 5, and 7].

3 Implications for the DWSG System

The basic theory that was developed in Part 2 will be used to make calculations specifically for the DWSG system. As the results are obtained, a discussion of their significance will be given.

3.1 What are the Issues?

The previous sections consisted of a careful consideration of the theoretical limits for infrared detectors. In this analysis, we found equations which give the spectral noise equivalent power and the spectral detectivity under BLIP conditions. The equations show how these figures of merit depend on the wavelength of monochromatic signal radiation.

When calculating BLIP $D^*(\lambda)$ or BLIP $NEP(\lambda)$, it is assumed that the wavelength of the monochromatic signal radiation is the same as the cutoff wavelength λ_c of the detector. This assumption is made simply because the ideal detector will have maximum response at the cutoff wavelength.⁹ However, when the DWSG is used to test a detector, it is exceedingly unlikely that the monochromatic laser radiation will just happen to be near the cutoff wavelength. Consequently, the figures of merit for BLIP observed under these special test conditions cannot have the same values as the usual figure of

⁸Apparently, spectral detectivity is the detectivity that is normally reported, although this is not always explicitly stated. The usual notation is D_λ^* , but the subscript is sometimes left out.

⁹In practice, λ_c is usually defined as the wavelength at half the peak responsivity (i.e., detectivity), but for most real detectors the difference between λ_c and the peak wavelength is quite small.

merits. Another big issue is the source of background photons. The previous sections assume a blackbody distribution when calculating the figure of merit, but as we shall see this will not always be appropriate in the DWSG. This difference will also change the values for the figures of merit at BLIP.

Will the detectors be in BLIP? Well, when testing IR devices, it usually is desirable to keep the background flux very low. The goal is to allow the testing to reveal the quality of the detector—that is, the device noise. If this is the case, then the detector will not be working under BLIP conditions. But, the DWSG is designed to test the detector under simulated mission conditions. This circumstance is very different than normal laboratory testing. If the designer of the IR detection system was able to minimize system and device noise, then it would seem likely that the detector is working very close to BLIP. Certainly this would be the goal of the system engineer who wishes to maximize performance.

The danger here is that when simulating the mission environment with a monochromatic laser beam significantly different from λ_c , the noise (as indicated by the figures of merit) will be artificially raised because of a change in the BLIP limit. There seems to be three different situations that can arise:

- The background flux from the lab is sufficient to put the detector into BLIP.
- The laser beam is used to “paint” a background according to mission expectations and the background flux from the room is negligible.
- The laser beam is used to “paint” a background according to mission expectations and the background flux from the room is *nonnegligible*.

Each of the following sections will be devoted to one of these cases.

3.2 Laboratory Background Flux Causes BLIP Conditions

This case is the simplest to deal with and perhaps the most unrealistic. The background flux is essentially a blackbody distribution as before, and we take the laser photons to contain nothing but signal information. Now the monochromatic signal flux used in all the equations from the theory sections is provided by the laser. The big difference, however, is that this laser is

operating at a wavelength different than the cutoff wavelength that is used in calculating the best possible performance of the detector. What we must do is find equations that will show how the performance under DWSG testing relates to the normal figures of merit.

We assume that the lab background flux which is large enough to put the detector into BLIP, is also significantly greater than the laser signal flux. Thus the conditions of Eqs. (9-10) are met. Under these conditions, the spectral noise equivalent power that will be observed is given by Eq. (14):

$$NEP(\lambda_l) = \frac{hc}{\lambda_l} \sqrt{\frac{2 \Delta f \phi_b^q(T)}{\eta}}. \quad (17)$$

Here we have put the subscript l on the wavelength to denote the signal as being generated by the laser. The normal situation, where the wavelength is λ_c , is simply given by:

$$NEP(\lambda_c) = \frac{hc}{\lambda_c} \sqrt{\frac{2 \Delta f \phi_b^q(T)}{\eta}}. \quad (18)$$

The relationship between these two figures of merit is easily seen to be:

$$NEP(\lambda_c) = \frac{\lambda_l}{\lambda_c} NEP(\lambda_l). \quad (19)$$

By using the definition of spectral detectivity, given in Eq. (15), it is easy to put the above relationship in terms of $D^*(\lambda)$:

$$D^*(\lambda_l) = \frac{\lambda_l}{\lambda_c} D^*(\lambda_c). \quad (20)$$

What do the last two equations mean? According to Eq. (19), if $\lambda_l/\lambda_c < 1$, then the laser will have to provide more power at λ_l than that required at λ_c to get the same signal-to-noise ratio. By the same token, if $\lambda_l/\lambda_c > 1$, then the laser will need to provide less power.¹⁰

If the laser wavelength is less than the cutoff wavelength, then the detector will behave as if its detectivity has been reduced by a factor of λ_l/λ_c . Another

¹⁰It is my understanding that occasionally the laser radiation from the DWSG is at a higher wavelength than the intended cutoff wavelength of the detector. This is accomplished by altering the temperature of the detector so it will be sensitive at the higher wavelength.

way of thinking about this correction is to consider the effect it has on Fig. 1. The line that represents the performance of an ideal photovoltaic detector operating at BLIP will be scaled down (see Fig. 2) by the same factor λ_l/λ_c . In other words, when testing with the DWSG the detector will have a different ideal limit.

In practice, detectors do not quite achieve the BLIP limit. Thus, if the limit is raised under testing it will have no effect. Lowering the limit will necessitate a correction if the test condition BLIP limit becomes lower than the detector's limit. The correction will not be quite as simple as Eq. (20) indicates because it is necessary to know the detector's actual limit under normal conditions beforehand.

3.3 Entire Incident Photon Flux Generated by the Laser

When using the DWSG to test an IR system in a mission environment, the laser will be used to "paint" the expected scene. Obviously, most of this scene will consist of what would normally be called background and, as discussed in Section 3.1, should put the detector near BLIP. Because the background is intentionally painted by the laser, this condition should be termed "simulated BLIP." The background photons responsible for simulated BLIP are not merely the normal photons with a blackbody-like distribution, but they include the monochromatic laser-generated photons.

We will now consider the case where virtually all photons incident on the detector are generated by the laser, and the flux from the lab is negligible. Such a situation is exactly the opposite of that expressed by Eqs. (9-10). Now we have

$$\phi_b^q(T) \ll \phi_s^q(\lambda_l). \quad (21)$$

Note that all the photons generated by the laser are legitimately considered part of the signal flux. After all, the detector has no way of distinguishing between those laser photons that represent the target and those that represent the background.

We are really working in the signal photon detection limit described in Section 2.2. Under these unusual circumstances, Eq. (5) gives the value for

the spectral noise equivalent power that would be observed:

$$NEP(\lambda_l) = \frac{2hc\Delta f}{\eta\lambda_l}. \quad (22)$$

The normal situation has a spectral noise equivalent power given by Eq. (18). A little algebra will show that these two equations are related in the following way:

$$NEP(\lambda_c) = \frac{\lambda_l}{\lambda_c} \sqrt{\frac{\eta\phi_b^q(T)}{2\Delta f}} NEP(\lambda_l). \quad (23)$$

Once again, Eq. (15) allows us to easily convert to spectral detectivity:

$$D^*(\lambda_l) = \frac{\lambda_l}{\lambda_c} \sqrt{\frac{\eta\phi_b^q(T)}{2\Delta f}} D^*(\lambda_c). \quad (24)$$

The interpretation of these two relationships is only slightly more difficult than in the last section. The ratio λ_l/λ_c plays the same role. But, due to the fundamental difference between the monochromatic laser photons and the usual blackbody background radiation, there is an additional, complicated correction term. This term depends on the quantum efficiency of the detector, the bandpass of the IR system, and the stray blackbody flux from the lab. The dependence on η and Δf is inconvenient because these variables will change for different detectors and systems. The dependence on $\phi_b^q(T)$, however, is unlikely to cause any difficulties because it should be a constant depending only on the DWSG.

In order to understand the implications of Eqs. (23–24) more fully, let us simplify things by considering a very bland scene. This simplification allows us to make a very rough order of magnitude calculation to get some feel for the minimum size of the new correction term. Assume a small quantum efficiency of 0.2. Assume the background flux density from the lab is the 10^{10} photons/cm²/sec measured for the baseline DWSG. A typical detector element has an area of 4×10^{-5} cm². Then, we find the flux is $\phi_b^q(T) = 4 \times 10^5$ photons/sec. An estimate of bandwidth is more difficult for this writer. It has been shown[8] that in theory the fastest possible response time of a typical HgCdTe detector is $\sim 0.2 \mu\text{sec}$. This speed sets an upper limit on the bandwidth of ~ 1 MHz.¹¹ Plugging these numbers into the radical of

¹¹For a system with an exponential decay time τ , the equivalent electrical bandwidth Δf is $1/4\tau$ (cf. Ref. 1, p. 115).

Eq. (24) should provide the smallest estimate of the minimum value the new correction term can have:

$$\sqrt{\frac{\eta\phi_b^q(T)}{2\Delta f}} \sim 0.2. \quad (25)$$

If λ_l/λ_c is slightly less than one, as will usually be the case, then the correction term will be almost an order of magnitude less than unity. At first glance this would indicate a significant reduction of the BLIP limit under DWSG testing. Thus, detectors would appear to have a detectivity that is artificially reduced even further than the reduction indicated under the conditions of the last section. However, the assumptions that went into the order of magnitude calculation of Eq. (25) were chosen to yield the extreme minimum correction factor. It is unlikely that this term would ever be so small. Although it is very possible that the lab background flux may eventually be reduced below the baseline levels, this change will, hopefully, still be swamped by the over-estimate of the bandwidth.

The author believes that the correction term will always be greater than unity. Consequently, under these conditions the BLIP limit will always be raised during testing with the DWSG. This result may seem surprising at first, but on closer inspection it is not unreasonable. A monochromatic source is providing the photons that under mission conditions would be generated by objects that emit a wide range of wavelengths. It seems physically reasonable to expect a simple "black and white" signal to have inherently less noise than a "multicolor" signal. If this is true, and the IR detector is limited only by BLIP, then the detector would output a cleaner signal under simulated tests than it will on the actual mission.

A higher BLIP limit, however, will not be a difficulty for the DWSG. Real detectors do not quite achieve the BLIP limit. Consequently, artificially reducing the fundamental noise due to the quantum nature of light should not enhance the detector's performance. The detector is dominated by noise from other sources and will continue to exhibit essentially the same detectivity even as the BLIP limit is raised.

3.4 Incident Flux Generated by Both Lab Background and Laser

When the background flux seen by the detector is generated both by the lab and the laser, then the situation is a combination of the cases discussed in the last two sections. Not surprisingly, things are now a bit more complex. Neither Eq. (21) nor Eq. (10) is valid. The background flux from the lab, and the laser generated signal photon flux, must be considered comparable:

$$\phi_b^q(T) \sim \phi_s^q(\lambda_l). \quad (26)$$

The following calculation will proceed in much the same way as the one in Section 2.3. As before, the signal current is simply

$$i_s = e\eta\phi_s^q(\lambda_l). \quad (27)$$

But now, the noise current which was given by Eq. (11) has an additional term. Thus,

$$i_n = e\sqrt{2\Delta f\eta[\phi_b^q(T) + \phi_s^q(\lambda_l)]}. \quad (28)$$

The signal-to-noise ratio is

$$S/N = \frac{i_s}{i_n} = \frac{\eta\phi_s^q(\lambda_l)}{\sqrt{2\Delta f\eta[\phi_b^q(T) + \phi_s^q(\lambda_l)]}}. \quad (29)$$

To find noise equivalent power, we need to set the signal-to-noise ratio equal to one. Then Eq. (29) can be solved for the signal flux. Multiplying the signal flux by the energy of a single photon gives the desired result:

$$NEP(\lambda_l) = \frac{hc\Delta f}{\lambda_l\eta} \left[1 + \sqrt{1 + \frac{2\eta}{\Delta f}\phi_b^q(T)} \right]. \quad (30)$$

Once again, we must compare this equation with the normal situation given by Eq. (18). After some algebra,

$$NEP(\lambda_c) = \frac{\lambda_l}{\lambda_c} \sqrt{\frac{2\eta}{\Delta f}} \left[\frac{\sqrt{\phi_b^q(T)}}{1 + \sqrt{1 + \frac{2\eta}{\Delta f}\phi_b^q(T)}} \right] NEP(\lambda_l). \quad (31)$$

Although complicated, this result seems reasonable. If the lab background flux is low enough so that $\phi_b^q(T) \ll \Delta f/2\eta$, then Eq. (31) easily reduces

to the same equation obtained in the last section for the all laser-photons case. Such a reduction is expected because all laser-photons implies a low lab background flux. On the other hand, if lab background flux dominates so that $\phi_b^q(T) \gg \Delta f/2\eta$, then Eq. (31) will reduce to Eq. (19). This result was obtained in Section 3.2 when all the background photons originated from the stray blackbody radiation in the lab.

This third case is probably the most realistic of the three considered. The first two situations are merely the limits of Eq. (31) at high $\phi_b^q(T)$ and low $\phi_b^q(T)$ compared to $\Delta f/2\eta$. In actual practice it would seem likely that the true value will not be at either extreme. Although the lab background flux will probably be relatively constant in the DWSG, the correction term will be different from one detector system to another if either the bandpass or quantum efficiency vary.

4 Conclusions and Recommendations

In this paper we have carefully considered whether the performance of an IR detector could change when it is illuminated by monochromatic radiation instead of the usual blackbody distribution. Such an analysis is important for understanding the effectiveness of the simulation testing which will be performed with the DWSG. It seems fair to conclude that the simulation may be affected because of changes in the fundamental detectivity limit.

Three cases have been identified. The first case occurs when all the background of the simulated scene is generated by stray radiation from the lab. Under this circumstance the BLIP limit will probably be reduced enough that the detector's performance is degraded. In the second case, the background flux from the lab is negligible. Then the BLIP limit is raised so the performance of the detector should be unaffected. The first two cases are really the extremes of the more general third case. Consequently, depending on the relative size of the lab background flux, the detector's measured performance may or may not be degraded.

At this point, there are several paths left to follow:

1. The third case should be carefully examined with numerical values appropriate for relevant IR detectors. We need to determine which parameter is the critical one and if the BLIP limit is likely to be raised

or lowered. Probably, a complete understanding of the bandpass Δf will be crucial.

2. If a detector is very close to the ideal limit, we need to confirm that its performance will not improve when that limit is raised.
3. In this paper it was assumed that the quantum efficiency is a constant when in practice it may vary by a factor of two. Will this wavelength dependence have a significant effect on the calculations reported here?
4. A measurement of the detectivity of a detector should be attempted in the DWSG and compared with the value found under standard conditions. Such an experiment should prove to be an important check on the validity of the present analysis.
5. Some consideration of the noise characteristics of a laser should be made. The assumption inherent in this paper is that the laser photon flux obeys a Poisson distribution. How valid is this assumption?

Hopefully, this paper has identified and correctly analyzed one of the major issues in the development of the DWSG. It will take, however, a significant amount of effort to apply the results to the actual data obtained with the DWSG.

5 Acknowledgements

I wish to thank Heard S. Lowry and P. David Elrod of Calspan Corporation for suggesting this line of research. Their pioneering efforts on the DWSG have stimulated my thinking in a number of ways. I am very grateful to my colleagues at Bethel College, Robert A. Carlsen and Thomas Greenlee. I had many fruitful conversations with Bob, and I appreciated the sound advice I received from him. Tom read a draft of this paper and asked good questions that helped me find a significant error. Once again, my wife Kim had the patience and provided the encouragement that I needed.

This work was made possible by a Research Initiation Grant from the Air Force Office of Scientific Research, Bolling AFB, DC.

6 References

1. Robert W. Boyd, *Radiometry and the Detection of Optical Radiation*, Wiley, New York, 1983, Chapters 7 and 8.
2. Eustace L. Dereniak and Devon G. Crowe, *Optical Radiation Detectors*, Wiley, New York, 1984, Chapter 2.
3. Robert H. Kingston, *Detection of Optical and Infrared Radiation*, Springer-Verlag, New York, 1978, Chapter 2.
4. Paul W. Kruse, "The Photon Detection Process," in *Optical and Infrared Detectors*, ed. Robert J. Keyes, 2nd ed., Springer-Verlag, New York, 1980.
5. Paul W. Kruse, Laurence D. McGlauchlin, and Richmond B. McQuistan, *Elements of Infrared Technology: Generation Transmission, and Detection*, Wiley, 1962, Chapter 9.
6. John D. Vincent, *Fundamentals of Infrared Detector Operation and Testing*, Wiley, 1990, Chapter 2.
7. S. F. Jacobs and M. Sargent III, *Infrared Physics* **10**, 233 (1970).
8. Brian P. Beecken, *Response of Infrared Detectors to Pulsed Radiation*, unpublished report, 1989.

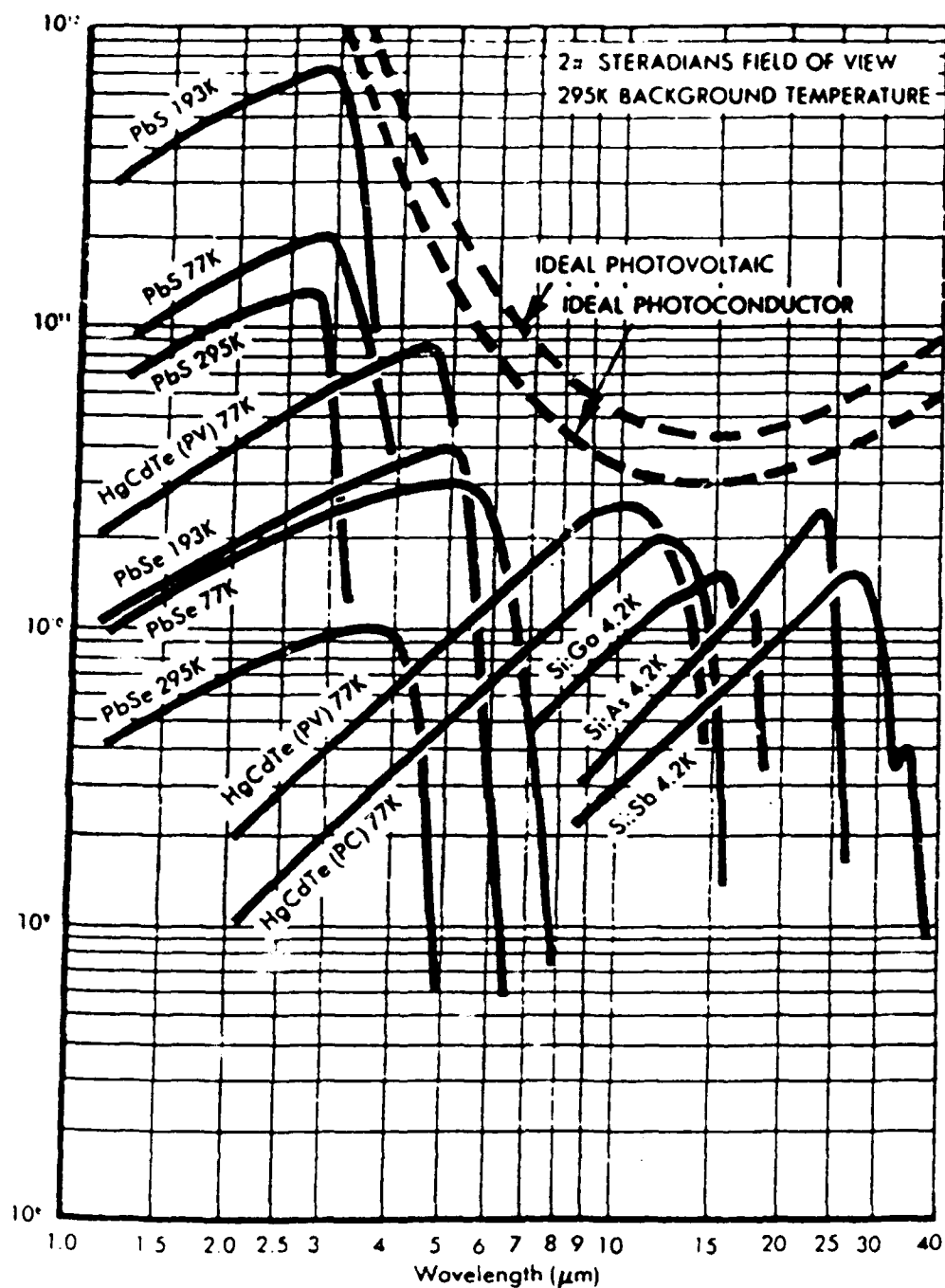


Figure 1: A reproduction of a portion of the SBRC *Infrared Wall Chart* which plots spectral detectivity as a function of wavelength.

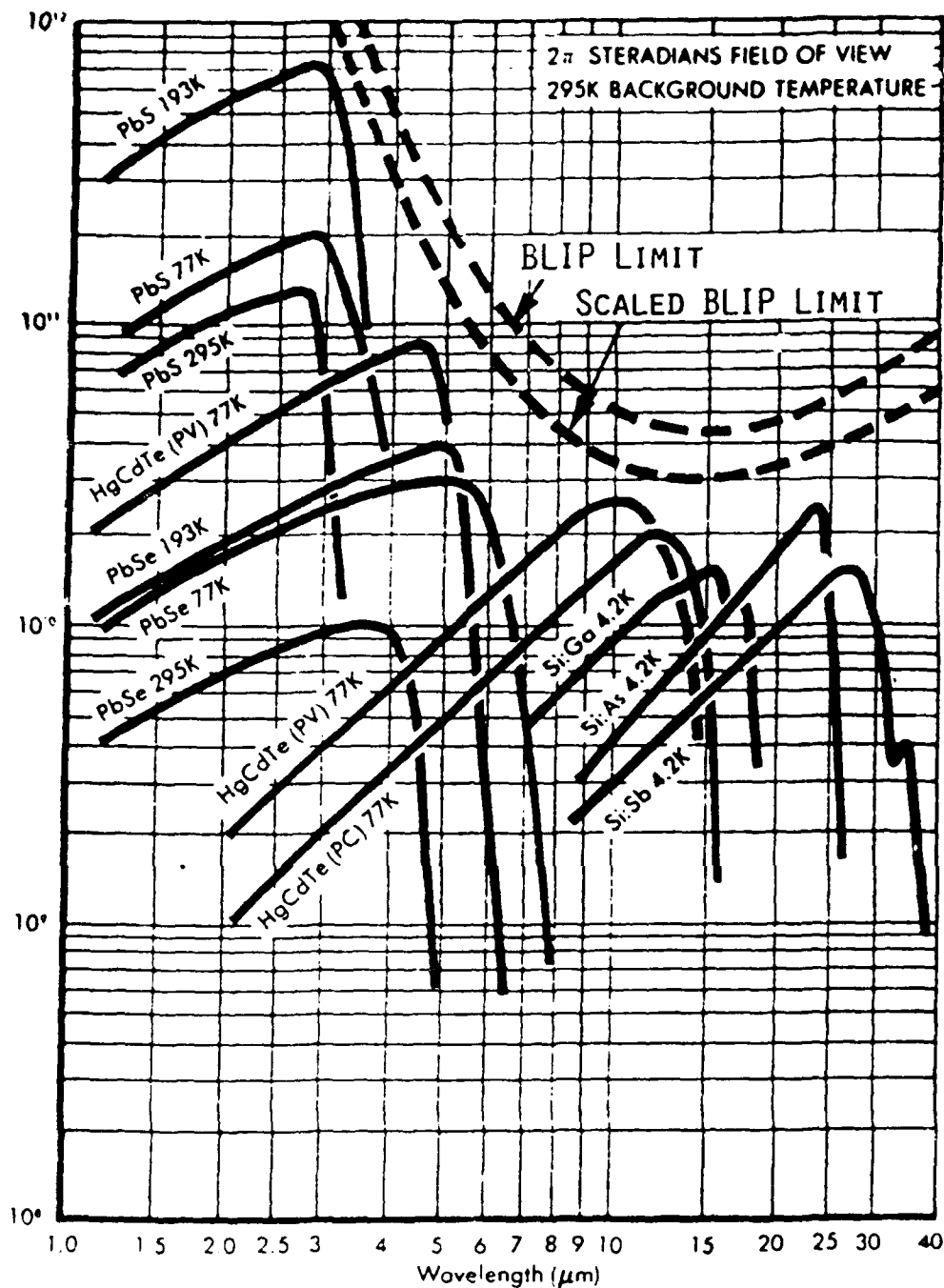


Figure 2: This variation of the SBRC *Infrared Wall Chart* shows the BLIP limit scaled downward by a factor of $\lambda_i/\lambda_c = 0.7$. The lower dashed line is the BLIP limit for a photovoltaic detector being tested in the DWSG system under the conditions described in Section 3.2. The upper dashed line is the same as in Figure 1.

1990 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by Universal Energy Systems, Inc.

FINAL REPORT

Sodium Fluorescence Studies for Application to
RDV of Hypersonic Flows

Prepared by:	Dr. Stephen H. Cobb
Academic Rank:	Assistant Professor
Department and	Department of Physics
University:	Murray State University
Research Location:	AF/DOTR
	Arnold Engineering Development Center
	Arnold AFS, Tennessee
USAF Researcher:	Carl Brasier
Date:	December 11, 1990
Contract No.:	F49620-88-C-0053/SB5881-0378

Abstract

Measurements of absorption line broadening and shift have been performed for sodium vapor with nitrogen as a perturber gas. In these studies, the sodium cell was maintained at temperatures of 200-350°C, while N₂ buffer gas pressures ranged from 10-600 Torr. Collapse of the hyperfine doublet of the Na ground state into a single peak was investigated as a function of probe laser intensity and buffer gas pressure. These studies are preliminary experiments in an effort to develop a Resonant Doppler Velocimetry technique for hypersonic flows at Arnold Engineering Development Center (AEDC).

Acknowledgements

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research. I also thank those at Arnold Engineering Development Center (AEDC) for allowing me to participate in this effort, and Universal Energy Systems for administration of this program.

I am particularly grateful to Mr. Carl Brasier at AEDC for conducting the experimental work. His support, advice, and experience made this effort most rewarding.

Introduction:

Resonant Doppler Velocimetry (RDV) is a non-intrusive diagnostic technique which uses laser-induced fluorescence from atoms or molecules in a flowing gas to measure the velocity, temperature, and pressure of the gas. RDV was proposed by R.B. Miles in 1975 (1), and first demonstrated in 1977 (2). The gas flow field is probed with a tunable dye laser directed so that it has a component in the direction of the flow. Since the atoms are moving, they exhibit a Doppler shifted peak absorption frequency. The fluorescence resulting from this shifted absorption may be compared to that of a stationary system, allowing the determination of the component of velocity in the direction of the beam.

Sodium atoms are often used as the fluorescing species in RDV experiments (3). Sodium atoms have two strong absorption line manifolds in the visible region. The wavelength of these lines is very near the peak output of an argon-ion pumped Rhodamine 6G dye laser. Also, sodium is found to occur naturally in some high temperature exhaust systems, eliminating the need to seed the flow via external means in experiments of this type.

The work described here is part of an effort to monitor and record sodium fluorescence features under a variety of temperatures, pressures, and perturber gas conditions. Analysis of the data will lead to a characterization and recognition of sodium fluorescence features for prospective test environments involving hypersonic flows.

Objective:

Sodium is well-known for its strong visible fluorescence features. Fluorescence studies of sodium to determine number densities, line broadening

parameters, and line shift parameters are well documented (4, 5, 6). However, in order to effectively implement the RDV method, it is crucial that the nature of the fluorescence profile in a stationary system be understood. With this goal in mind, studies have been initiated to observe the position and shape of sodium fluorescence features as a function of perturber gas constituent, temperature, and pressure.

Apparatus:

The excitation and detection system for observing sodium fluorescence is shown in schematic form in Figure 1. A Spectra-Physics Model 171 Argon-ion laser was used to pump a Coherent 699-21 ring dye laser. Using Rhodamine 6G dye, the ring laser was able to deliver 1W of single frequency radiation peaking at approximately 590 nm. The ring dye laser is capable of scanning a 30 GHz frequency range in 0.25 sec. A Fabry-Perot interferometer monitored the laser line to check single mode frequency stability. The laser was directed into a temperature controlled fluorescence cell (Comstock heat pipe) which could be evacuated to the millitorr range. Transmitted intensity was monitored with a silicon photodiode, and fluorescence was detected at 90° to the beam direction with an RCA 8644 photomultiplier tube with S-20 response. The detector output was collected by a LeCroy 9400 digital oscilloscope, triggered synchronously with the scanning electronics of the ring laser. Fluorescence plots were recorded on the scope face and dumped to an HP7475A plotter.

Experiment:

Sodium vapor in a heated cell was excited by a continuous wave ring dye

laser, tuned 4-30 GHz across the sodium D₁ resonance line. This resonance line is at a wavelength 589.6 nm, corresponding to the 3²S_{1/2} to 3²P_{1/2} transition in sodium (Figure 2). Absorption and fluorescence profiles were monitored for sodium cell temperatures ranging from 200-350°C. This corresponds to sodium number densities in the range approximately 10¹³ - 10¹⁵ atoms/cm³. These profiles were recorded in order to measure the resonance lineshift and line broadening present for different pressures of perturber gas. In all experiments, the perturber gas introduced into the sodium cell was nitrogen.

McCartan and Farr have shown the broadening and shift of the sodium D lines to depend linearly on perturber gas pressure up to about 0.8 atm (7). These studies were conducted using Ar rather than N₂ as the perturber gas. Jongerius et. al. (5) have measured the collisional broadening and shift of the Na-D lines in a vapor cell at approximately 460K containing Ar, N₂ and H₂ as perturber gases. According to their results, in a vapor cell with N₂ at 460K the broadening coefficient is 2.74 x 10⁻²⁰ cm⁻¹/atom/cm³. Ideally, the number density is related to temperature and pressure by the following:

$$n = 2.7 \times 10^{19} \left(\frac{P}{P_{\text{atm}}} \right) \left(\frac{300}{T(^{\circ}\text{K})} \right) \quad (1)$$

So that at 460K we expect a broadening of

$$(2.74 \times 10^{-20} \text{ cm}^{-1}) (2.3 \times 10^{16} \cdot P(\text{Torr})) \quad (2)$$

or 18.7 MHz per Torr change in perturber pressure.

Experiments performed by AEDC personnel with N₂ perturber pressures ranging from 10-600 Torr were performed in a Na vapor cell at 526K. These studies were found to yield a broadening parameter of 19.9 ± 9 MHz per Torr pressure change.

Jongerius also reports an observed pressure shift coefficient for N₂ perturber gas as $0.82 \times 10^{-20} \text{ cm}^{-1}/\text{atom}/\text{cm}^3$. This shift is toward lower frequencies, and would correspond to a frequency shift of 5.55 MHz per Torr pressure change. Miles (1) has put forth a pressure-temperature relationship for frequency shift, which for N₂ can be expressed as

$$\Delta f = -6.76 \text{ Ghz} \cdot P(\text{atm}) \cdot \text{SQRT}(300\text{K}/T) \quad (3)$$

At 460K, this results is $\Delta f = 7.2$ MHz per Torr change in perturber pressure. AEDC experiments at 526K yield red shifted frequencies for Na in N₂ of approximately 7.7 MHz per Torr change in perturber pressure.

The set point temperature and pressure of the sodium vapor cell used in the AEDC experiments proved difficult to maintain, thus the data obtained were not as repeatable as desired. It was noted that in many trials, the maximum

observed absorption of the excitation beam could reach 40%. This allowed the possibility of significant lineshape distortion due to radiation trapping. It was observed that for any given pressure, it was possible to see changes in the fluorescence lineshape by altering the intensity of the probe beam. This phenomenon, discussed by Walkup et. al (8), was investigated for its probable importance in accurately determining the peak frequency of absorption and fluorescence in RDV experiments.

Sodium has a ground state hyperfine splitting of 1.77 GHz (Figure 3) which can easily be resolved by the instruments used in this investigation. The absorption lineshape for sodium vapor is expected to show the superposition of two Voigt profiles separated by 1.77 GHz and weighted in intensity by the degeneracy factors of the hyperfine levels, 5 to 3. Walkup observed the collapse of this hyperfine doublet into a single peak due to laser-induced optical pumping at laser intensities far below radiative saturation levels. Conditions were calculated which showed this lineshape distortion to occur when the product of laser power and buffer gas pressure exceeded 0.1 mW Torr. Walkup's work was conducted with 10 Torr of Xenon as the perturber gas.

This altering of lineshape was also observed in AEDC experiments. In order to predict pressure and intensity regimes where lineshape distortion might occur, calculations were performed using the method of Walkup (8). The absorption rate per unit volume is expressed as

$$A = \frac{n_1^0 R_1 [D + (\alpha_{12} + \alpha_{21}) R_2] + n_2^0 R_2 [D + (\alpha_{12} + \alpha_{21}) R_1]}{D + \alpha_{12} R_1 + \alpha_{21} R_2} \quad (4)$$

where the R factors represent the stimulated absorption rates and are Voigt profiles.

$$R_1 = .5 \gamma_s^2 (I/I_s) \int_{-\infty}^{\infty} dv \frac{e^{-v^2/v_t^2}}{\sqrt{\pi} v_t} \frac{\gamma}{\gamma^2 + (\omega - kv - \omega_1)^2} \quad (5)$$

Here, the spontaneous decay rate $\gamma_s = 6.3 \times 10^7/\text{sec}$; the saturation intensity $I_s = 12\text{mW/cm}^2$; propagation constant $k = 2\pi/\lambda$; the Lorentzian HWHM $\gamma = 0.5(\gamma_s + bp)$, where b is the pressure broadening coefficient, and $v_t = (2kT/m)^{1/2}$. In equation 4, the α 's are branching ratios and are chosen to be statistical ($\alpha_{12} = .625$ and $\alpha_{21} = .375$), as are the equilibrium ground state populations n_1^0, n_2^0 . Values for diffusion rate D and coefficient of pressure broadening b were taken from sources by given Walkup.

Figures 4–8 show the results of these calculations. In Figure 4, the overlapping Voigt profiles are denoted by asterisks while the solid line shows

the resulting absorption lineshape. In figures 5-8, the pressure is held constant while the laser intensity is increased, and the resulting lineshape profile is shown. Note that when the pressure intensity product reaches ~ 0.1 mW Torr, there is only one maximum evident, and this peak becomes stationary after the intensity reaches 0.01 mW. The fact that this peak location shifts to a final frequency alleviates fear that the absorption profile is in continuous motion and hence can't be used as a reference in RDV experiments. In fact, as the peak narrows with increased laser intensity, peak locations become more easily measurable. These calculations were repeated for pressures of 10, 50 and 200 Torr with similar results. It should be noted that increasing laser intensity can result in observation of Lamb dips in absorption profiles, which show up as dips or depressions at the peak of the absorption curve.

Conclusions:

More experiments are necessary for a complete understanding of sodium absorption and fluorescence in the presence of buffer gases at AEDC. Efforts should be made to increase the pressure and temperature stability of the test environment, and experiments should be carried out at low probe laser intensities to observe absorption lineshapes and shifts without saturation effects or optical pumping.

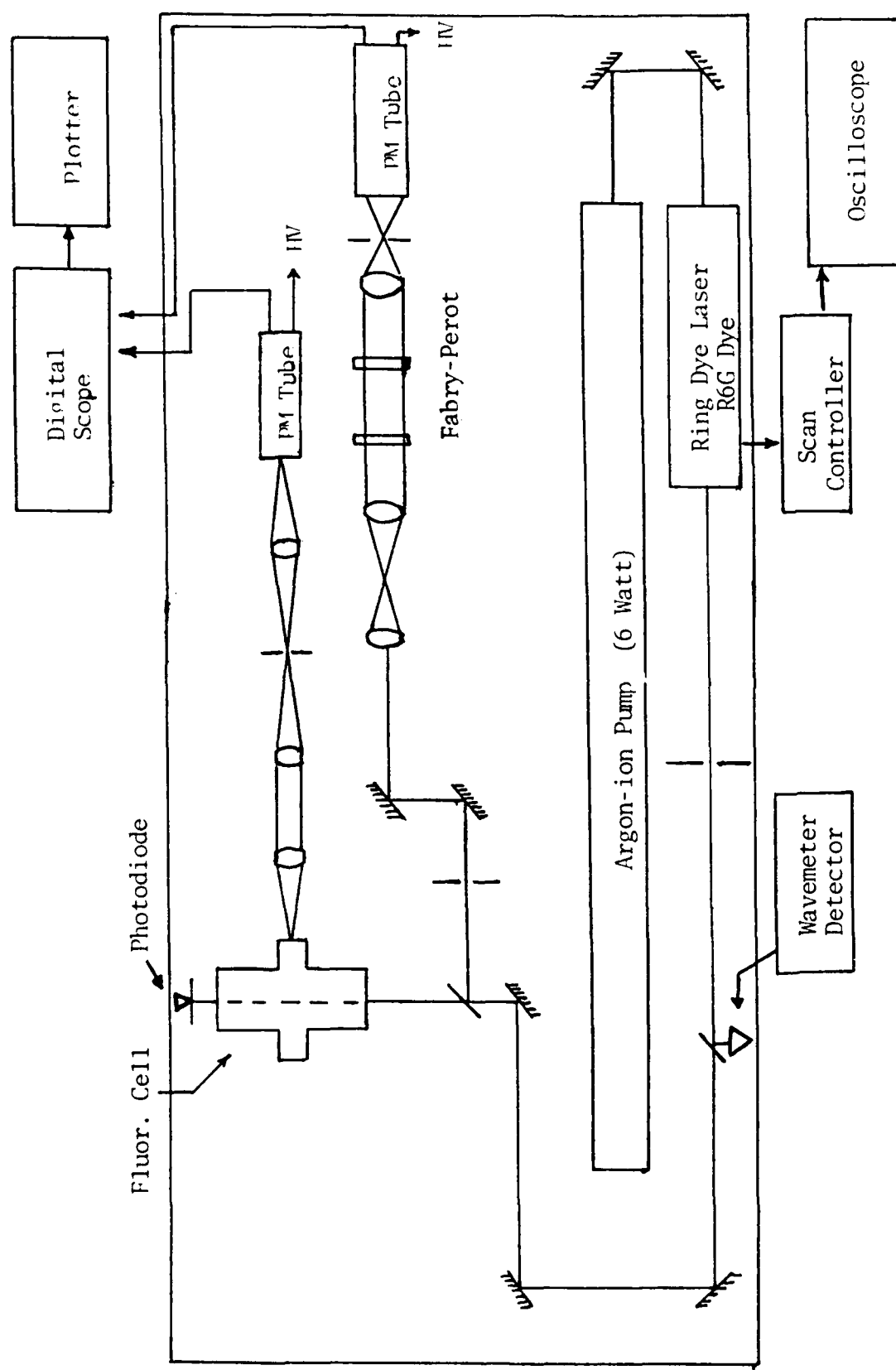


Figure 1. Schematic of Experimental Equipment

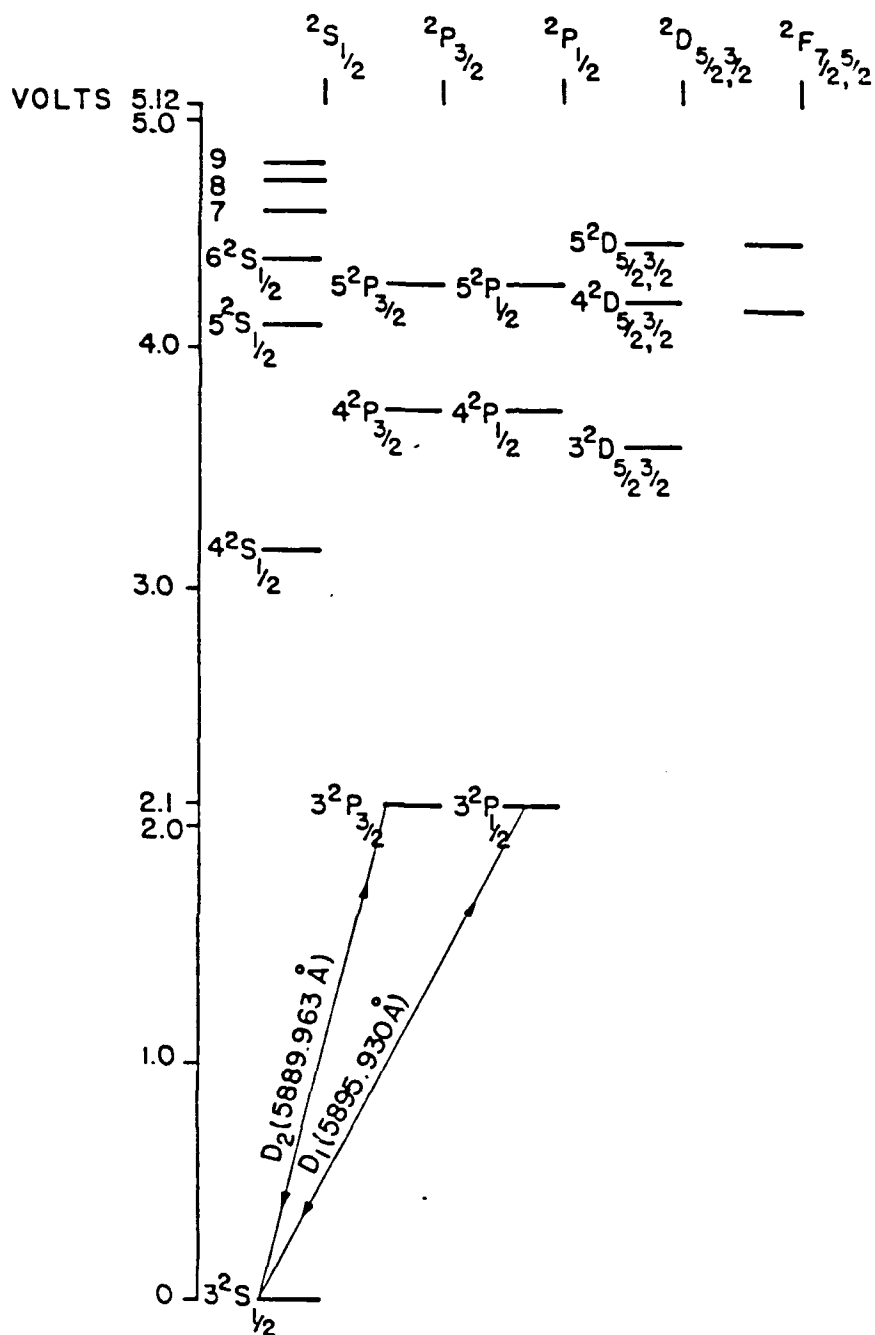


Figure 2. Energy levels of Sodium atom (3).

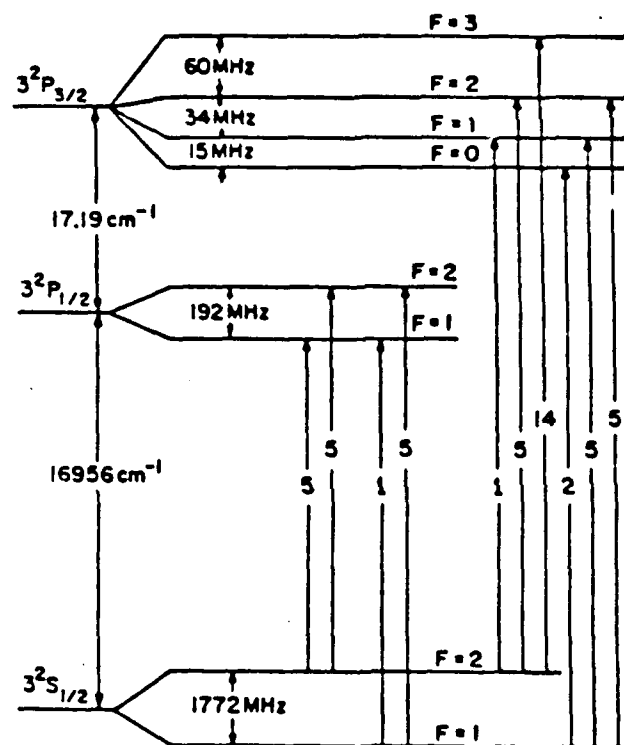


Figure 3. Hyperfine splitting of energy levels in the Na-D lines (4).

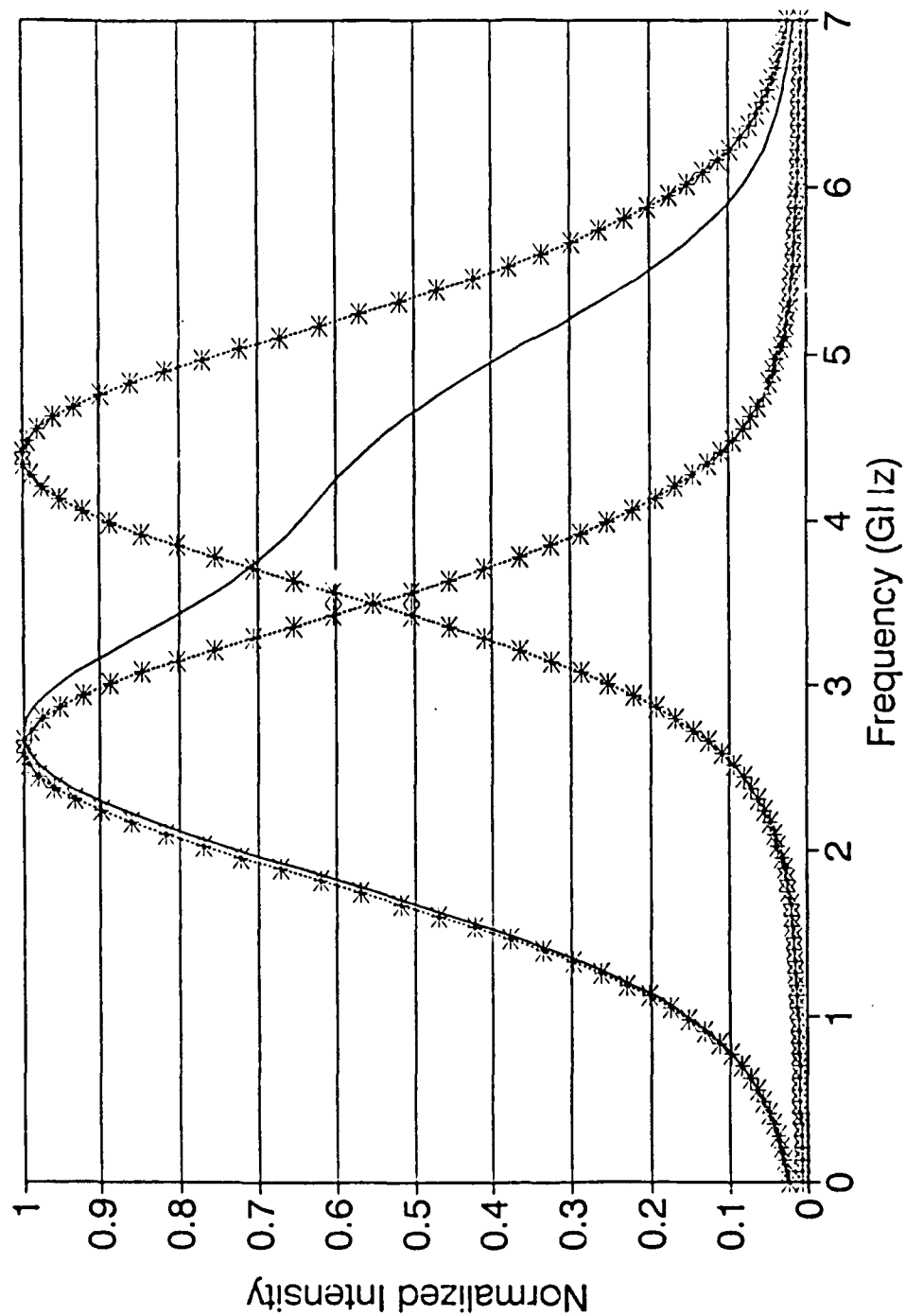


Figure 4. Fluorescence intensity for the Na-D₁ line (solid curve). Asterisks represent the two Voigt profiles which make up the solid curve. N₂ pressure is 100 Torr, laser intensity is 1E-6 mW.

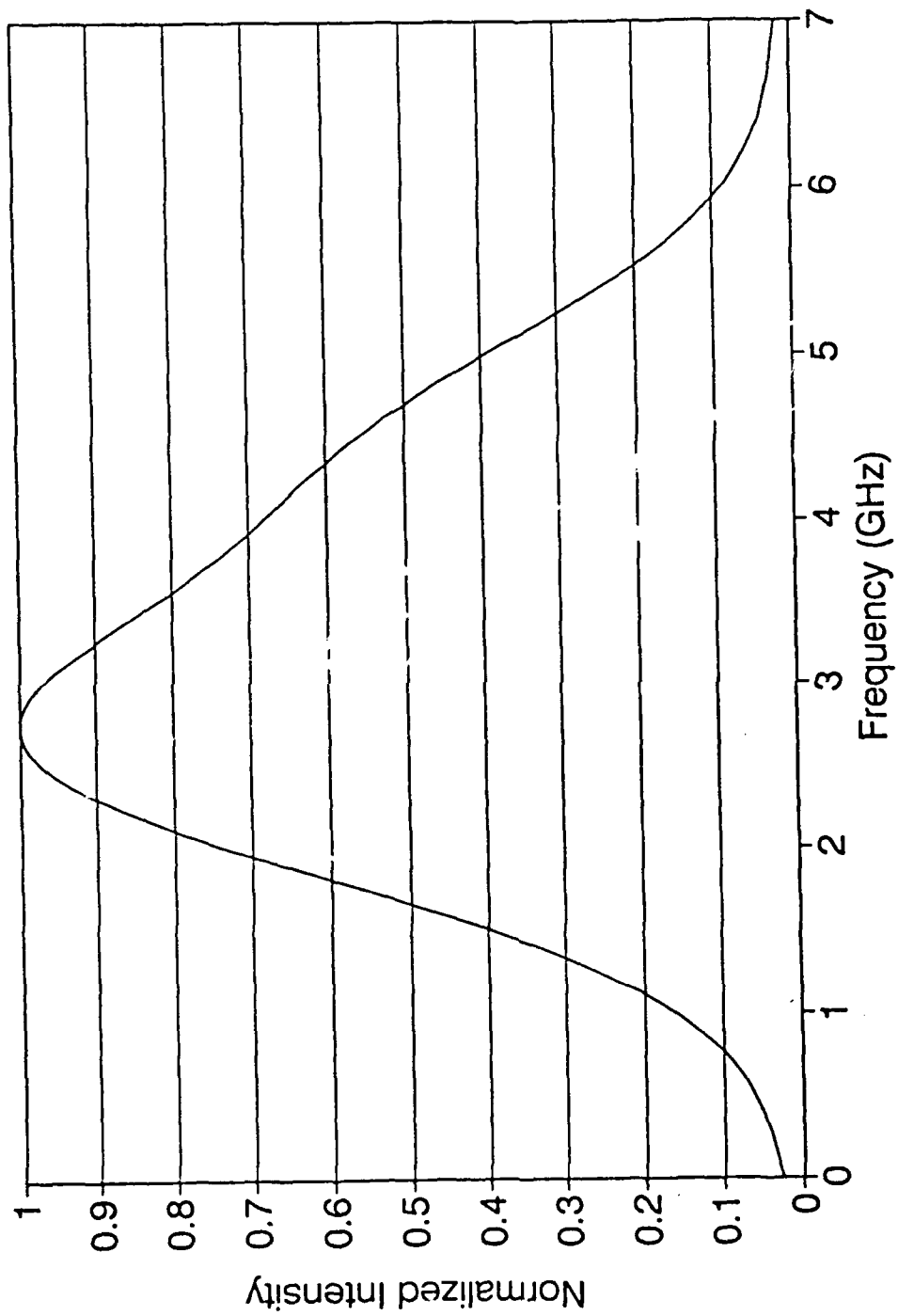


Figure 5. Fluorescence intensity for the Na-D₁ line. N₂ pressure is 100 Torr, laser intensity is 1E-4 mW.

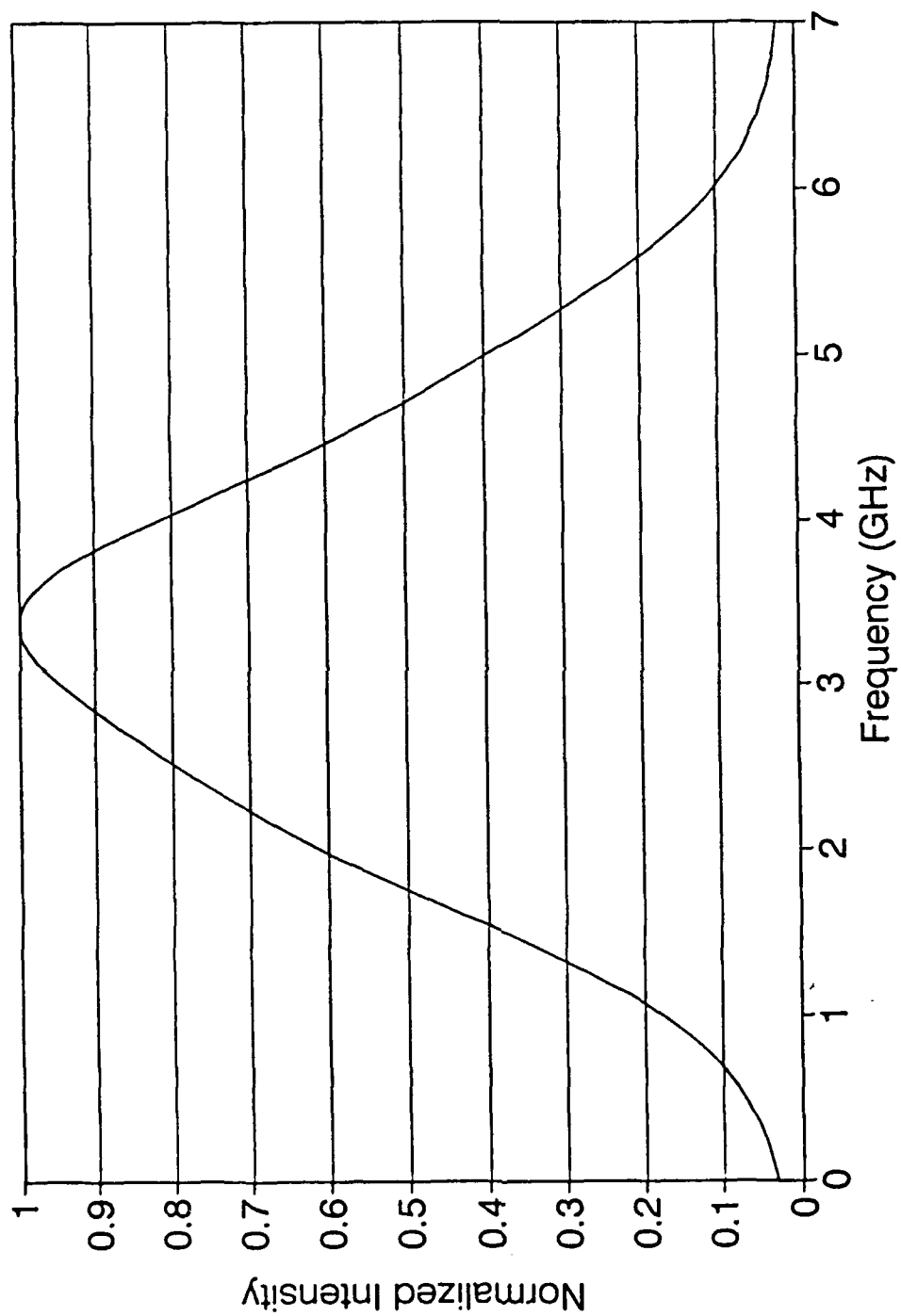


Figure 6. Fluorescence intensity for the Na-D₁ line. N₂ pressure is 100 Torr, laser intensity is 1E-3 mW.

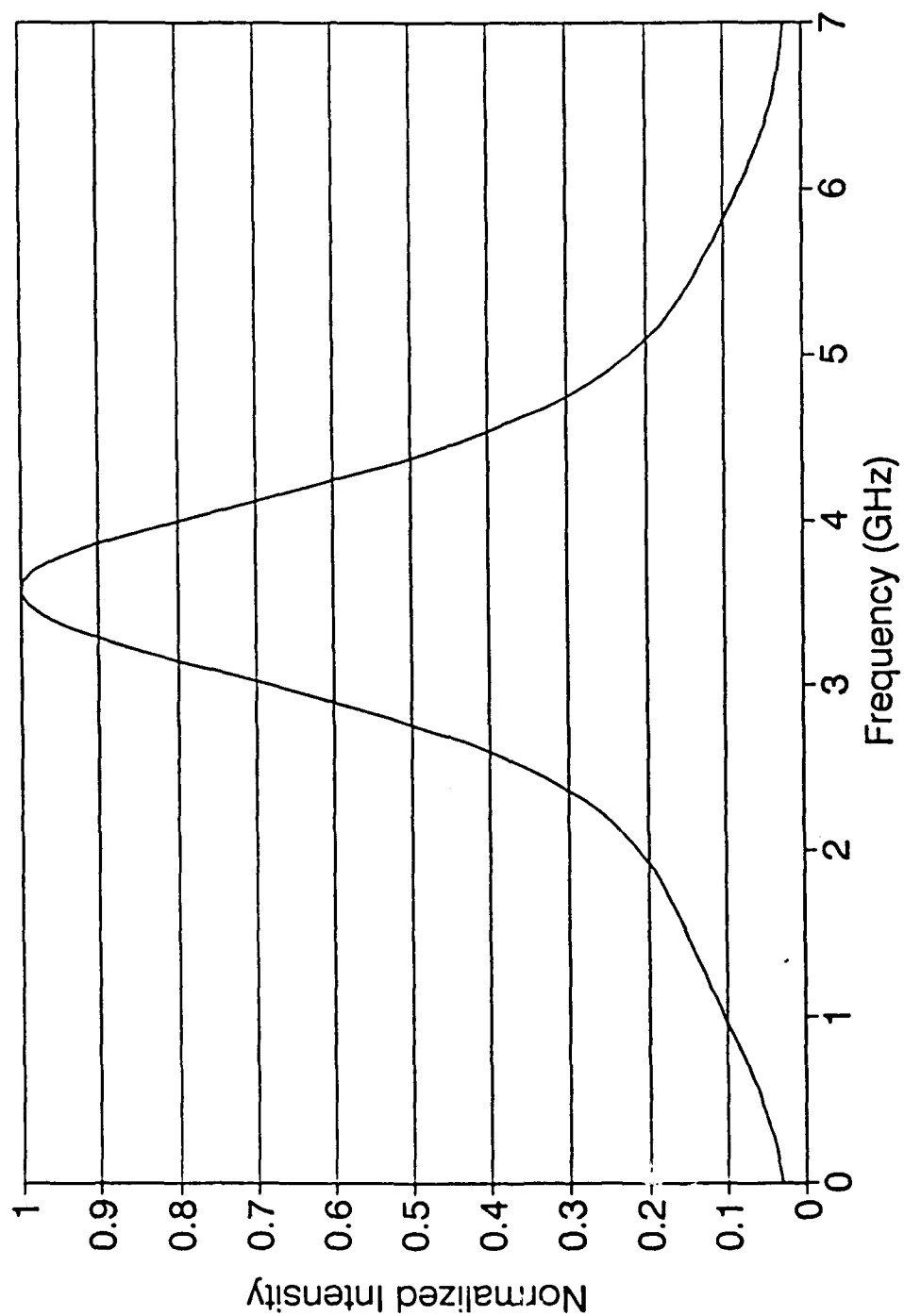


Figure 7. Fluorescence intensity for the Na-D₁ line. N₂ pressure is 100 Torr, laser intensity is 1E-2 mW.

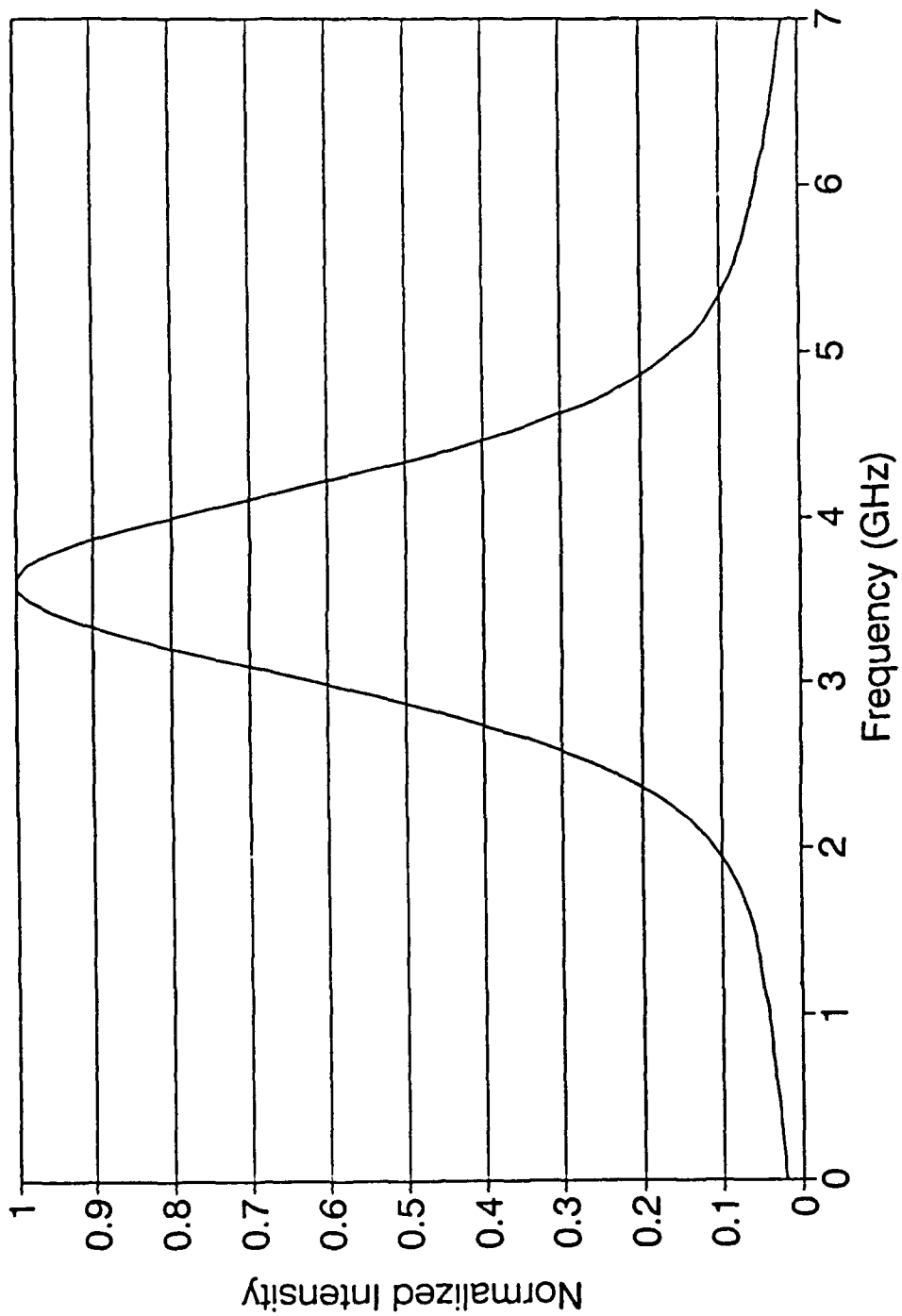


Figure 8. Fluorescence intensity for the Na-D₁ line. N₂ pressure is 100 Torr, laser intensity is 1E-1 mW.

References

1. R.B. Miles, Phys. of Fluids 18, 751 (1977).
2. R.B. Miles, E. Udd, M. Zimmermann, Appl. Phys. Lett, 32, 5 (1978).
3. S. Cheng, "Resonant Doppler Velocimetry in Supersonic Nitrogen Flow",
Ph.D. Thesis, Princeton Univ., Princeton, NJ, 1982.
4. W.M. Fairbank, T.W. Hansch, A.L. Schawlow, JOSA 65, 199(1975).
5. M.J. Jongerius, A.R.D. Van Bergen, T. Hollander, C.T.J. Alkemade, J. Quant.
Spect. Rad. Trans.25, 1(1981).
6. J.L. Lemaire, J.L. Chotin, F. Rostas, J. Phys. B. 19, 1913(1986).
7. D.G. McCartan, J.M. Farr, J. Phys. B. 9, 985(1976).
8. R. Walkup, A. Spielfiedel, W.D. Phillips, D.E. Pritchard, Phys Rev A 23,
1896(1981).

Report # 13
210-10MG-086
Prof. John Francis
Report Not Publishable At This Time

Report # 14
210-10MG-134
Prof. Orlando Hankins
Report Not Publishable

FINAL REPORT


**AN EXPERIMENTAL APPROACH FOR THE DESIGN OF A
MIXER FOR AN ARC HEATER**

for

The Air Force Office of Scientific Research
(Contract No. F49620-88-C-0053/SB5881-0378)

November 30, 1990

Submitted by



Lang Wah Lee
Professor of Mechanical Engineering
University of Wisconsin-Platteville
Platteville, Wisconsin

ABSTRACT

An experimental study on crossflow jet mixers was undertaken to evaluate the effect of various similarity parameters on mixing results. The tests were carried out in three mixers of similar shape but different sizes. Two common fluids, air and water, were used as the test mediums. In the air mixer, heated air jets were injected and mixed with a primary flow of air at ambient temperature; and in the water mixer, jets of salt brine were mixed with a primary flow of fresh water. The scalar fields (temperature or concentration) in the mixing chambers under various operating conditions were measured and the measured data afforded an assessment on the relative importance of the similarity parameters on mixing. Based on such results, several parameters pertinent to mixing operation were identified. It appears that the mass transfer process in a water mixer does not duplicate the heat transfer process in an air mixer. Thus the validity of using mass transfer process in water mixers to simulate the heat transfer process in air mixers is opened to question.

ACKNOWLEDGEMENT

I wish to thank the Air Force System Command and the Air Force Office of Scientific Research for giving me the opportunity to carry out this research. The support I received from the Universal Energy system is also greatly appreciated.

The work would not be possible without the consultation and arrangement from the personnels in AEDC, particularly from F.L. Shope and Carlos Tirres. In carrying out the research at the University of Wisconsin at Platteville, I received help and encouragement from Ross McDonald, Dean of College of Engineering, Richard Strunk, Chairman of Mechanical Engineering, and Chris Lind, Director of Sponsored Programs. The help from two of my students, Messrs Jeff Baltes and Jeff Gafner, and from John Abing, technician in Mechanical Engineering, is invaluable in ensuing the completion of the project. Finally but not leastly, I wish to thank Jerry Lolwing and Mary Duewer for typing the report.

NOMENCLATURE

a	constant for jet trajectory
b	constant for jet trajectory
\bar{c}	time average concentration at a spatial point
c_{j0}	concentration of salt brine before mixing
c_0	concentration of salt in primary fluid before mixing
d	diameter of the mixing chamber
D	diameter of the mixing chamber
I_s	intensity of segregation
J	momentum flux ratio
L	length of the far field
L_s	length of segregation
Pr	Prandtl number
Re_j	Reynolds number for the jet fluid
Re_0	Reynolds number for the primary fluid
Ri	Richardson number
S	pitch between injection ports
Sc	Schmidt number
St	Stanton number
t	time
t_c	time constant of mixing
t_R	resident time
T	temperature
T_{j0}	initial temperature of the jet fluid
T_0	initial temperature of the primary fluid
V_j	jet velocity

V_o velocity of the primary fluid
 x axial distance from the plane of injection
 x_o length of the near field
 z vertical distance from the injection port.

Greek Alphabet

E dissipation rate of turbulent energy
 θ dimensionless temperature
 ν kinematic viscosity
 ρ_j density of jet fluid
 ρ_o density of primary fluid
 σ standard deviation of concentration
 τ dimensionless time
 ϕ dimensionless concentration

I. INTRODUCTION

The idea of attaching a jet mixer to an arc heater was conceived at the Arnold Engineering Development Center (AEDC) to upgrade a facility for ground simulation of hypersonic flow in the re-entry flight of space vehicles. In the mixer, the jets are issued in a direction normal to the primary flow. The development of such a mixer was perceived as primarily an empirical task because a reliable analytical method for the mixing process has not been fully developed. On the other hand, the environment of extremely high temperature (more than 5000 K) and pressure (more than 150 atm) in the prototype mixer is very difficult and costly to duplicate in a pilot test facility; it is desirable to simulate the actual process with common fluids such as air or water at moderate temperature and pressure. In particular, AEDC has available a water tunnel test facility and it is desirable to perform test with that facility. The cost of the testing can be cut down even further if the heat transfer process can be modeled by a mass transfer process where salt brine is injected and mixed with the primary flow of fresh water in a water mixer. The validity of such an approach needs to be examined on the basis of similarity law for the process in jet mixers. A complete set of similarity parameters for a jet mixer, to the author's knowledge, has not been published in the literature. Previous works were often confined to two types of studies, namely, (1) the study of jet trajectory for either a thermal plume (Wright, 1977) or an isothermal jet (Moussa et al., 1977); and (2), the study of mixing

in a pipe mixer (Forney et al., 1979).

In the summer of 1989, the author participated in the USAF-UES Summer Faculty Research Program and his task was to identify the similarity parameters for the mixing process to facilitate the preparation of an experimental program for the design of the mixer. As the result of that work, a set of dimensionless similarity parameters was developed (Lee, 1989) and this analysis was further updated in a later work to include the effect of heat transfer process between the jet stream and the primary fluid. This set includes a total of nine dimensionless parameters of which two are related to the geometry of the mixer and the other seven are derived from equations governing the mixing process. The research work also found that it was not feasible to satisfy all the conditions imposed by the seven parameters unless the two fluids being mixed in the pilot mixer are exactly the same as those fluids in the prototype mixer. However, such a stringent requirement can be relaxed if some of the seven parameters are found to be of secondary importance to the mixing process. The relative importance of the seven parameters can be found only through experiments.

This work is a continuation of the research started in the summer of 1989. The goals are to rank the relative importance of the seven parameters through an experimental program and to check whether the process in a water mixer duplicates that in a gas mixer. To this end three mixers of similar shape but different sizes were designed and built as the testing facility. Of the

three mixers, two were water mixers where salt brine was injected into a primary stream of fresh water and the resulting concentration field was measured; the third was a gas mixer where heated air was injected to a flow of ambient air and the resulting temperature field was measured. By varying the experimental conditions the effect of each parameter on mixing can be isolated for evaluation. The results from such experiments, together with those from other investigators, afford an assessment of the effect of each parameters. The comparison between the concentration field in the water mixers and the temperature field of the gas mixer illustrates whether mixing in a water mixer can simulate the mixing in a gas mixer.

This report includes a brief review of previous works, followed by a description of the test apparatus and experimental procedures, the presentation of experimental results, discussion and conclusions.

II. OBJECTIVES

1. To design and construct three model mixers for evaluating similarity criteria for the mixing process in jet mixers.
2. To compare the relative importance of the similarity parameters and to identify the ones pertinent to the mixing process.
3. To examine the validity of simulating the heat transfer process in the prototype mixer with the mass transfer process in a water mixer.
4. To recommend an experimental approach to facilitate the design of the prototype mixer.

III. BASIC THEORY

According to the analysis of Lee (1989), the flow in the mixer can be divided into two zones, the near field and the far field. The near field is located between the injector and the point where the jets impinge on one another. In the near field, the identity of each jet remains and the mixing is accomplished through entrainment. The rest of the mixer downstream of the near field is called the far field where all the jets lose their identity and the mixing is accomplished by turbulent diffusion. Corresponding to these two mixing mechanisms, two sets of similarity parameters were deduced by Lee (1989). A summary of this work is given in what follows.

A. The Near Field: The similarity parameters were derived from the method of physical similitude (Kline, 1986). In this method, the dimensionless parameters were deduced by taking the ratio of

the individual terms of the conservation equations. Thus the conservation equations of mass, momentum and energy for a jet in cross flow were invoked and the method, led to the following dimensionless parameters

$$\frac{T - T_{j0}}{T_o - T_{j0}} = f \left(\frac{\rho_j V_j}{\rho_o V_o}, St, Ri, Re_j, Re_o \right)$$

A more general relationship should also include the effect of geometrical factors, which consist of the diameter ratio between the injector and the mixer, and the pitch to diameter ratio (a parameter related to the number of injection ports). The inclusion of these two parameters yields

$$\frac{T - T_{j0}}{T_o - T_{j0}} = f \left(\frac{\rho_j V_j}{\rho_o V_o}, St, Ri, Re_j, Re_o, \frac{d}{D}, \frac{S}{D} \right) \quad (1)$$

The dimensionless parameters shown in Eq.(1) are more general than those given by Moussa et al (1977) (whose study was confined to non-buoyant flow) and those provided by Wright (1977) and Forney et al (1982) (which concerned mainly with mixing in a pipe mixer). In fact, it can be shown that the dimensionless parameters deduced from these investigations are only a subset of those shown in Eq.(1).

As is well known from dimensional analysis, the functional relationship in Eq.(1) can only be obtained empirically and to achieve such a goal would require extensive testing. The goal of this study, however, is to identify the parameters pertinent to the mixing process and to compare the mixing process in air and water mixers.

B. The Far Field. In turbulent mixing theory, the degree of mixing is specified by the intensity of segregation I_s which represents the difference between the ingredients being mixed. Corrsin (1957, 1964) showed that in isotropic turbulence the intensity of segregation would decay exponentially with time, i.e.

$$I_s = \frac{\text{rms value of temperature fluctuation at time } t}{\text{initial rms value of temperature fluctuation}} = e^{-t/t_c} \quad (2)$$

where t_c is the time constant of mixing. The value of t_c can be evaluated from the following equations.

i, for gas $Pr \text{ (or } Sc) \leq 1$

$$t_c = \left(\frac{5}{\pi}\right)^{1/2} \frac{2}{3-Pr^2} \left(\frac{L_s^2}{\epsilon}\right)^{1/2} \quad (3)$$

ii, for liquid $Pr \text{ (or } Sc) \gg 1$

$$t_c = \frac{1}{2} \left[3 \left(\frac{5}{\pi}\right)^{1/2} \left(\frac{L_s}{\epsilon}\right)^{1/2} + \left(\frac{\nu}{\epsilon}\right)^{1/2} \ln Pr \right] \quad (4)$$

By normalizing equation (2), Lee (1989) obtained the following relations:

$$I_s = e^{-t_R/t_c} \quad (5)$$

where t_R is the resident time of the fluid in the mixer

$t_R = L/V_0$, and τ is a dimensionless time

Thus the dimensionless parameter in the far field is the time ratio t_R/t_c . As can be seen from Eqs.(3) and (4), the time constant t_c is dependent on L_s (the scale of segregation) which is in turn dependent on the result of mixing in the near field. Thus the degree of mixing in the far field is related to all the

the degree of mixing in the far field is related to all the parameters specified in Eq. (1), in addition to the resident time in the mixer.

Combining the dimensionless parameters for the near field and the far field, one obtains

$$\frac{T - T_o}{T_\infty - T_o} = F \left(\frac{\rho_j}{\rho_o}, \frac{V_j}{V_o}, St, Ri, Re_j, Re_o, \frac{d}{D}, \frac{S}{D}, t_R/t_c \right) \quad (6)$$

For mass transfer process, the temperature should be replaced by concentration and St should be the Stanton number for mass transfer.

IV. TEST APPARATUS AND EXPERIMENTAL PROCEDURE

A. Test Apparatus

1. The mixers

Three mixers, two for aqueous flow and the third for gaseous flow, were designed and built at the University of Wisconsin-Platteville. The three mixers are similar in shape, but different in size (the diameter of the mixing chamber for the two water mixers is respectively 4 and 6 inches, and is 8 " for the gas mixer). A schematic drawing of the three mixers is shown in Fig. 1. As can be seen from the drawing, a mixer consists of three sections, i.e. the head, the middle and the end sections. The head section includes the inlet for the primary fluid, a diffuser, a flow straightener, and a settling chamber. The middle section consists of a collar and the mixing chamber. Four injecting nozzles are mounted with even spacing around the circumference of

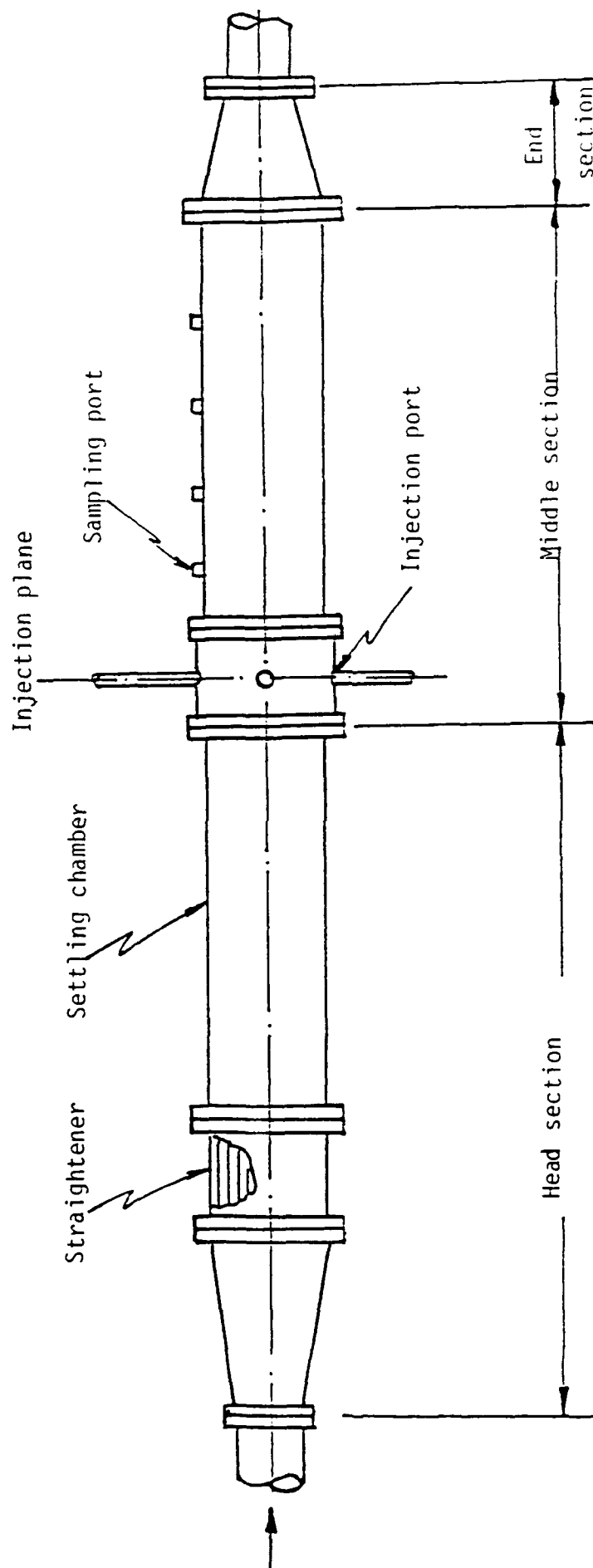


Fig. 1 Schematic of the mixer

the collar on a plane normal to the axis of the mixer. The four nozzles can be either closed or opened so that the number of the injection ports can be varied. By changing the size of the nozzle one may also vary the jet diameters. The end section is a convergent flow passage connecting to the exit of the mixer. Holes are drilled along the wall of the mixing chamber either for taking samples (as for water mixers) or for installing thermocouple probes or velocity probes (as for the gas mixer). The holes are located at $0.375 D$, $0.75 D$, $1.5 D$, $3 D$, $4.5 D$, and $6 D$ from the injector plane.

The flow system for the two water mixers is shown in Fig. 2. The primary fluid (fresh water) was supplied from a constant head tank (600 gallons, 63 ft head above ground) and a recirculation pump. The injected fluid (salt brine) was pre-mixed in a brine tank and then delivered to the injection nozzles by a centrifugal pump via a manifold. The flow rates of the primary fluid and the injected fluid were measured respectively by a turbine flow meter and four orifice meters. The photographs of the two water mixers are shown in Fig. 3a and 3b.

The flow system (Fig. 4) for the gas mixer consists of a radial flow fan ($5'' H_2O$ head, 1000 cfm) which is used to deliver primary flow of air at room temperature ($21^\circ C$) to the mixing chamber. The injected air is heated by an electric heater (1.3 KW) and delivered to the injection nozzle by a blower ($21'' H_2O$, 100 cfm). Orifice meters are used to measure the flow rates of both the primary and the injected fluids.

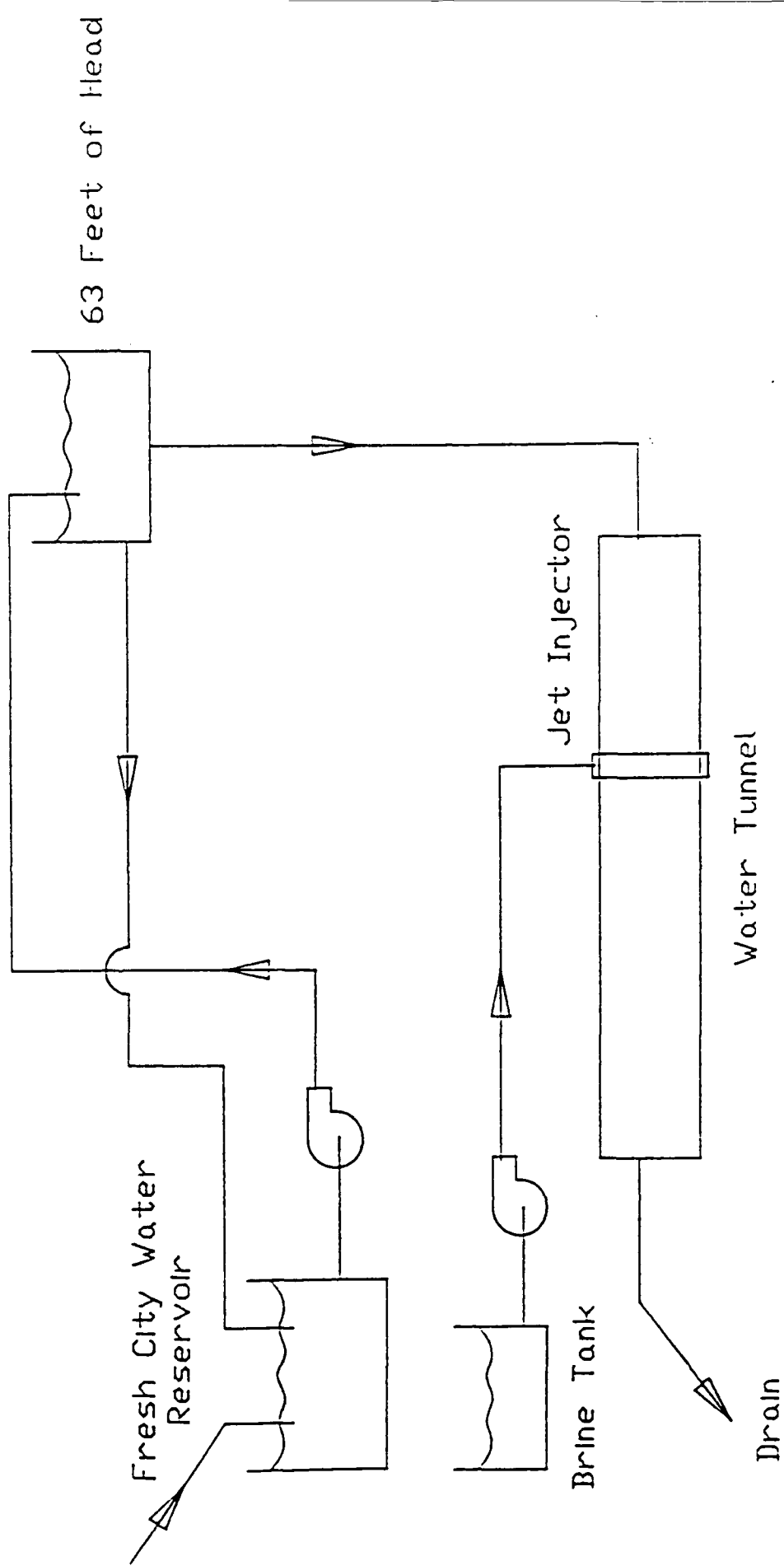


Fig. 2 Schematic of the water flow system



Fig. 3a The 4" water mixer

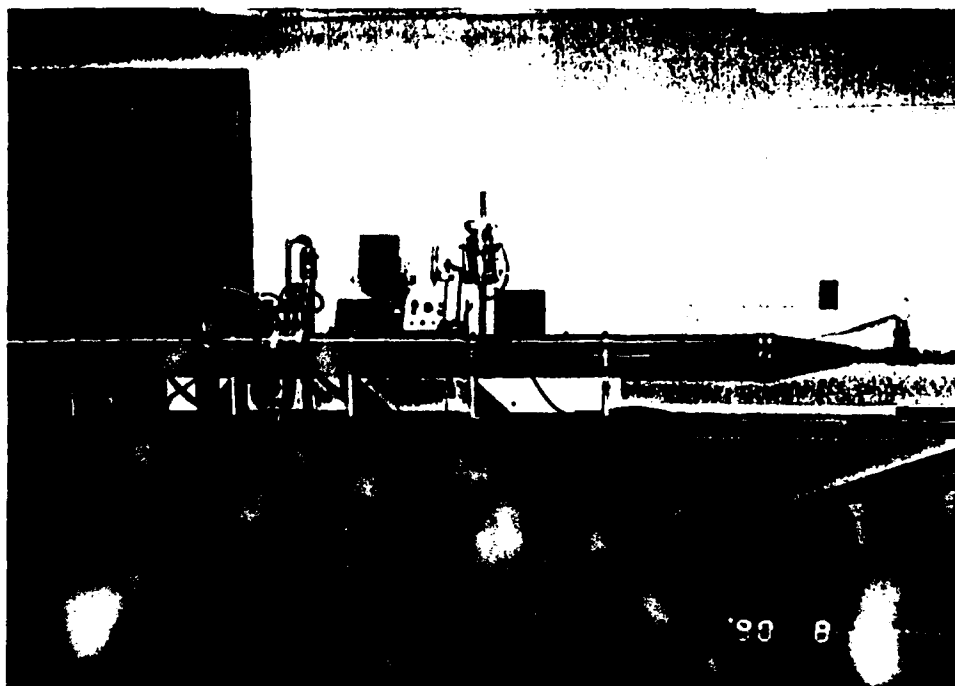


Fig. 3b The 6" water mixer

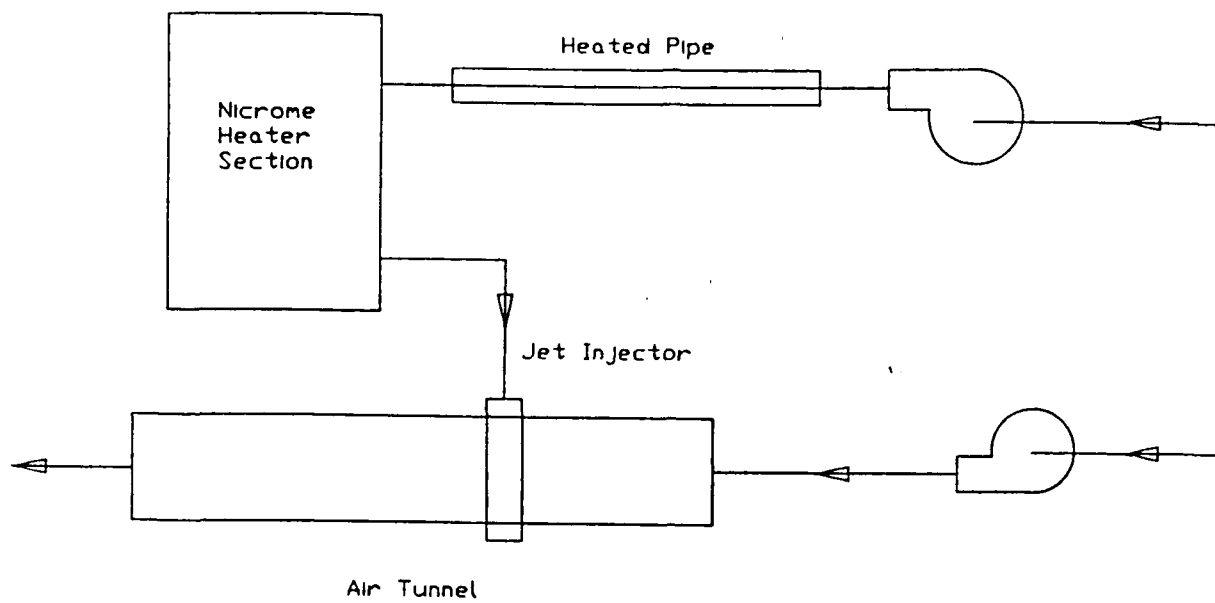


Fig. 4a Air flow system for the air mixer



Fig. 4b The air mixer system

2. Measurement Techniques

The concentration field in the two water mixers was obtained by taking samples (with hypodermic needles) from the sampling holes. The sampling location can be varied radially by traversing the tip of the needle, axially by inserting the needle in various sampling holes along the mixing chamber, and angularly by rotating the injection collar. The disturbance caused by the hypodermic needle was negligible because the size of the needle was very small (1/32" O.D.) and the rate of drawing was very slow (0.5 cc/sec). The sample (approximately 30cc) was then placed in a test tube and the concentration was measured by a conductivity meter (model CDB-70, Omega Co) whose accuracy was 0.5% of the concentration level being measured. The calibration of the meter indicated linear response over the range of measurement.

The temperature field in the gas mixer was measured with copper-constantan thermocouples. Nine thermocouples (AWG No. 32) were fixed to a rake over a distance of 4 inches. The rake was then inserted to the mixing chamber through the holes on the walls. The signal from the thermocouples was then recorded by a computer-data-acquisition system at a sample rate of 2000 Hz. Fifty readings were taken in each sample and the average temperatures were calculated through a software package.

The velocity field around the jet was measured with a five-hole direction probe.

B. Limits of the Measurement Technique

The measurement technique mentioned above can yield only the time average of the scalar field. From the measurement the average value over a cross section can be taken and the deviation from the average can also be deduced. The uniformity of the scalar field is an indicator of the mixedness in the mixer. However, the deviation of the scalar field from the average is different from the concept of intensity of segregation mentioned in Eq. (2) which is the rms value of the fluctuation measured at a spatial point with respect to time. Thus the method employed in this experiment cannot be used to verify the relations specified by Eq. (5).

C. Experimental Schedule

With the three mixers of different sizes operating with two different test mediums, one may examine the effect of the dimensionless parameters in Eq. (6) on mixing by varying the density ratio, the velocity ratio, the flow rate, the diameter of the nozzle and the number of injection ports. Comparison between the scalar fields in the air mixer and the water mixers can illustrate whether test on water mixer can simulate the test on air mixer.

The experimental matrix for the air mixer and the two water mixers is listed in Tables 1,2, and 3.

Table 1. Experimental matrix for the gas mixer (Primary fluid:
ambient air at 21°C)

Number of jets	Velocity ratio	Reynold number	Diameter ratio	Temperature of jet fluid (°C)
1	3.75	29500	1/16	97
1	3.75	88000	1/16	96
1	5	88000	1/16	97
1	5	29500	1/16	90
1	1.8	88000	1/8	111
1	5	88000	1/8	115
2	2	88000	1/16	98
2	3.75	88000	1/16	99
2	5	29500	1/16	98
2	5	50000	1/16	100
2	5	88000	1/16	100
2	6	88000	1/16	111
2	7	88000	1/16	115
4	3.75	88000	1/16	86
4	5	88000	1/16	84
4	5	50000	1/16	88
4	5	29500	1/16	86
4	7	88000	1/16	82

Table 2. Experimental matrix for water mixer #1
(Dia. = 4; primary fluid: fresh water from water main)

Number of jets	V_j/V	Re	d/D	Concentration of jet fluid C_j (%)
1	4	12000	1/16	7.5
1	4	16000	1/16	7.5
1	4	16000	1/16	15
1	4	16000	1/16	1.8
1	5	16000	1/16	7.5
4	3	16000	1/16	7.5
4	4	16000	1/16	7.5
4	5	16000	1/16	7.5

Table 3. Experimental matrix for water mixer #2
(Dia. = 6"; primary fluid: fresh water from water main)

Number of jets	V_j/V	Re	d/D	Concentration of jet fluid C_j (%)
1	3.75	30000	1/16	7.5
1	4	16000	1/16	7.5
1	5	30000	1/16	7.5
2	4	30000	1/16	7.5
2	5	30000	1/16	7.5

D. Indicator of Mixedness.

In this work, the degree of mixing (or mixedness) is defined as the standard deviation σ of the local time average concentration (i.e., a spatial point in a cross section) with respect to the average concentration over the entire cross section. The scalar field (either the temperature field or the concentration field) was first measured and the results were converted to the following dimensionless forms:

$$\text{a. temperature parameter} \quad \theta = \frac{T - T_o}{T_{j_o} - T_o} \quad (7)$$

$$\text{b. concentration parameter} \quad \phi = \frac{\bar{C} - C_o}{C_{j_o} - C_o} \quad (8)$$

According to Shope (1989) the standard deviation σ at the exit of the mixer should be less than 5%. The distance required by the fluid stream to attain such a degree of mixing will be called the mixing distance. Incidentally, this is also the minimum length of the mixer. Such a length is not measured in this work because the primary goal of the project is to examine whether it is feasible to

use a water mixer to simulate the mixing in an air mixer.

The standard deviation σ , used in this work to denote mixedness, is different from the intensity of segregation I_s (shown in Eq. (2)) which is used in chemical engineering practice. According to Brodkey (1967), the intensity of segregation I_s is ratio of the rms fluctuating value to the initial rms fluctuating value. To measure I_s , one must record the instantaneous scalar quantity over a certain period of time. The sampling method used in the water mixer could not yield such measurements. Thus the method employed in this work cannot be used as a direct verification of Eq. (5). However, since both I_s and σ decay along the direction of flow in the mixer, a measurement of σ can probably be regarded as an indicator of I_s , and thus indirectly reflects the effect of the time ratio t_r/t_c specified in Eq. (5). It must also be pointed out that the determination of t_c is rather difficult due to the lack of reliable methods to ascertain the values of L_s and ϵ in Eqs. (3) and (4). (McKelvey et al. 1975).

V. RESULTS

The effects of the dimensionless parameters specified in Eq. (6) on mixing are examined by examining their effects on the measured scalar fields. Prior to the presentation of such results, one must first have the knowledge concerning the extent of the near field and the far field, and this information can be obtained if the velocity trajectory of the jet is known. The velocity trajectory measurement will thus be given before the presentation

of the scalar field measurements.

A. Velocity Trajectory of the Jets.

There are several definitions with regards to jet trajectories. Here the jet trajectory is defined as the locus of the maximum velocity in the plane of symmetry, in a manner similar to that used by Kamotani et al. (1972). The intersection point between the trajectories of two opposed jets is considered as the center of coalescence and the distance between the center of the injection port and the point of coalescence is defined as the extent of the near field. Such a definition of the near field is only approximate because the impingement of the jets would cover a certain area instead of at a point (due to the finite dimension in jet width and thickness).

The measurement was carried out in the air mixer and the velocity vector field was obtained by using a five-hole directional probe. The result is shown in Fig. 5 along side the trajectories predicted by the following empirical equation developed by Kamotani et al. (1972):

$$z/d = a (x/d)^b \quad (9)$$

In Eq. (9), z and x represent respectively the penetration and the longitudinal movement of the jet; a and b are empirical constants. The comparison shown in Fig. 5 demonstrates that the measured trajectories are in reasonable agreement with Eq. (9) and further measurement does not seem to be necessary. Although the equation is for predicting the trajectory of a single jet in cross flow, it can be used for predicting the trajectory of multi-jet as well

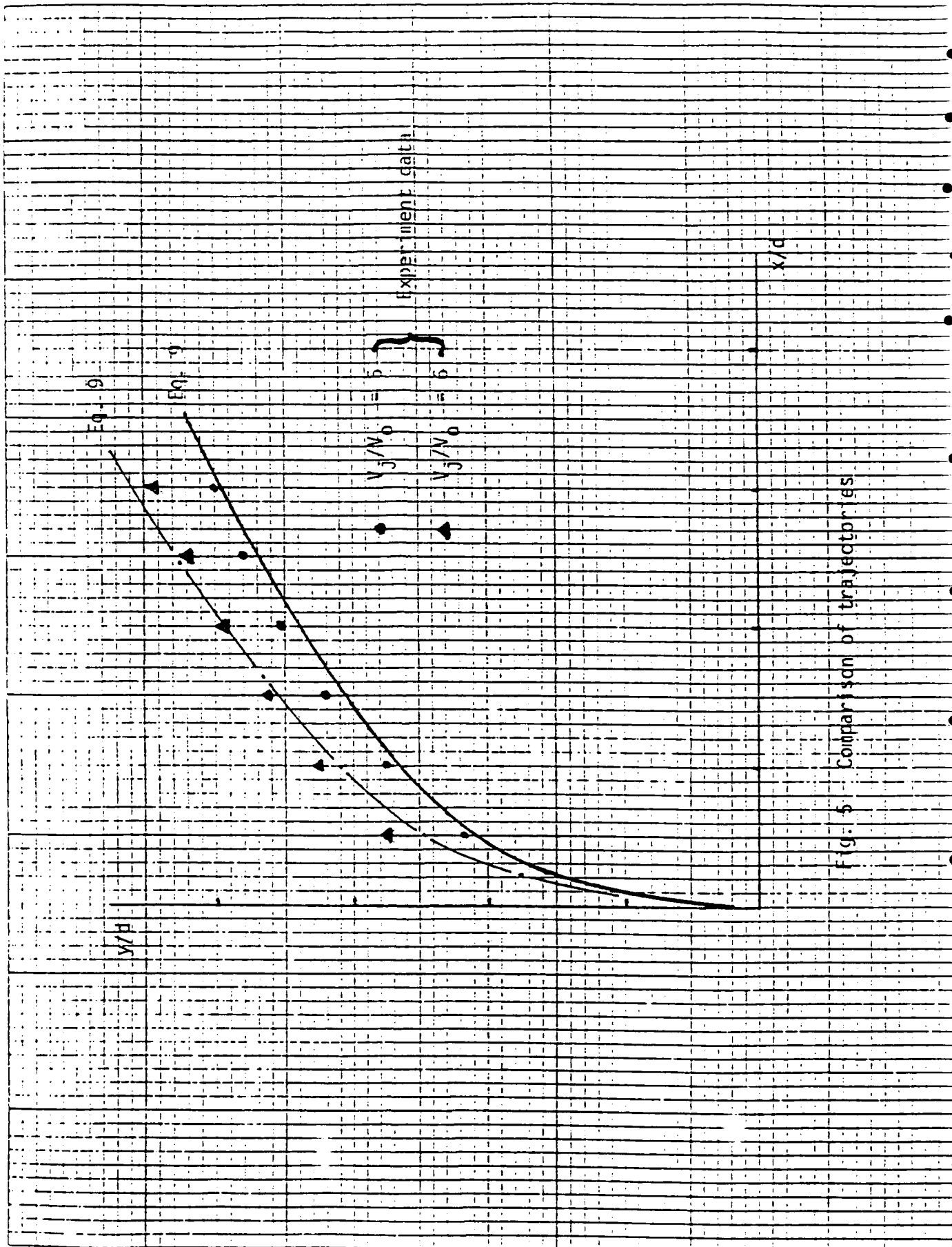


Fig. 5. Comparison of trajectories

(Kamotani, 1974). Thus Eq. (9) will be used in this work to predict the jet trajectory and the point of coalescence, and the result is shown in Table 4.

Table 4 Location of the point of coalescence at various velocity ratios

V_j/V_o	x_o/d
4	10.17
5	6.2
6	4.13
7	2.63

In Table 4, x_o is the axial distance between the center of the injection port and the point of coalescence. As expected, the value of x_o decreases as the velocity ratio increases.

B. Effect of Number of Injection Ports

Effect of number of injection ports on mixing was examined in both the gas mixer and the aqueous mixers. The presentation of the result will mainly focus on the experiments carried out in the gas mixer because they are more complete. The mean temperature profiles were plotted in contour maps for the case of a single jet, two diametrically opposed jets and four diametrically opposed jets. The velocity ratio in this plot is equal to 5.

The contour map of a single jet is shown in Fig. 6. It shows that the contour exhibits a characteristic kidney shape. The high temperature zone in the map corresponds to the location of a vortex which was also observed by other investigators. As the jet proceeds downstream, the spread and rise of the jet can clearly be seen by comparing the three maps in Fig. 6.

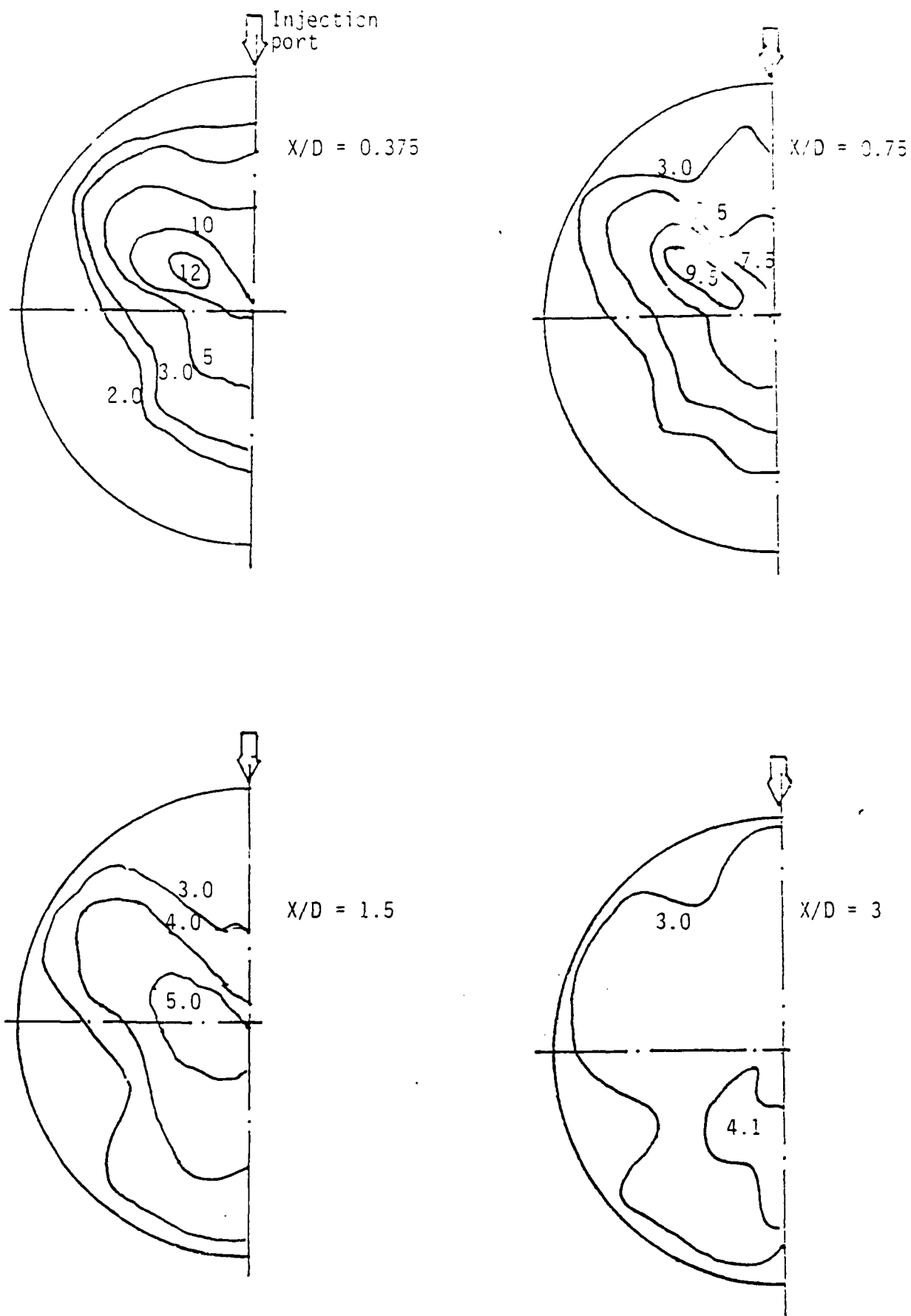


Fig. 6 Contour maps ($\theta = \text{constant}$) for a single jet air mixer
 ($V_j/V_o = 5$, $d = 0.75''$)

For the case of double jets, the contour maps of the temperature field are shown in Fig. 7. At the location of $x/d = 6$ (near the point of coalescence as shown in Table 4), the presence of the vortex is still discernible. After the two opposed jets impinge on each other, the two pairs of high temperature zone merge into one pair (only one vortex is shown) and spread sideways in a direction normal to the plane containing the jet center line.

Fig. 8 shows the contour maps for four jets (forming two pairs of opposed jets) at various axial locations. The two high temperature zones in each jet disappear, probably due to mutual interference from neighboring jets. The sideways spread observed in Fig. 6 is also hindered by the other pair of jets and the spread after the impingement is observed only in the direction of 45° with respect to the plane containing the center line of the jet.

The standard deviation of the temperature field at two cross sectional positions for these three cases is shown in Table 5.

Table 5 Standard deviation σ at two cross sections

	σ (at $x/d=50$)	σ (at $x/d=100$)
One jet	55%	
Two jets	36%	25%
Four jets	24%	15%

Table 5 indicates that mixedness increases with the number of injection ports. However, even with four jets operating, the standard deviation is still higher than the targeted 5% at a location 100 jet-diameter away from the injection port.

Tests being carried out in water mixers show the same tendency as observed in the gas mixer. However, the spread and standard

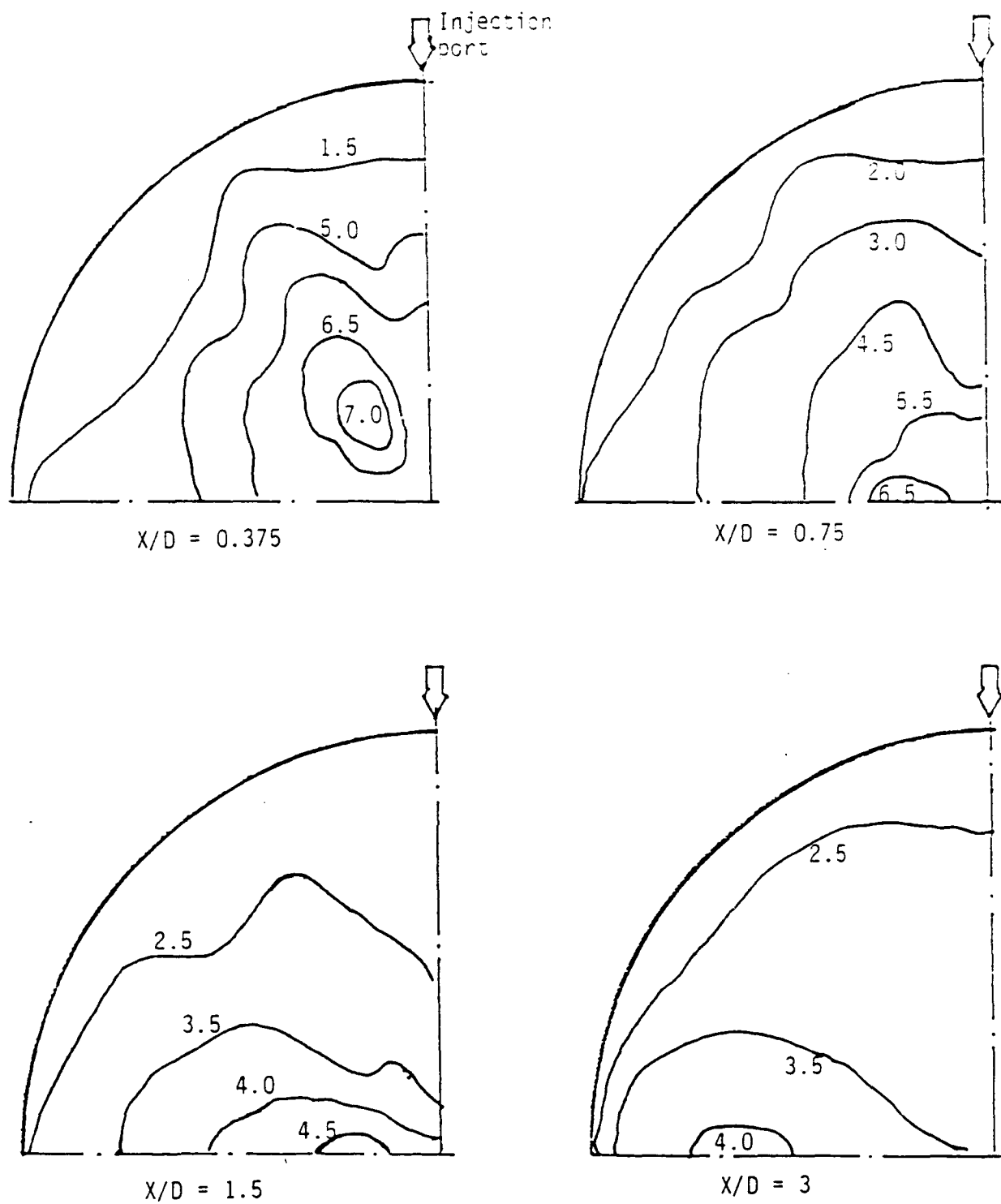


Fig. 7 Contour maps ($\theta\%$ = constant) for two-jet air mixer
($V_j/V_o = 5$, $d = 0.5''$)

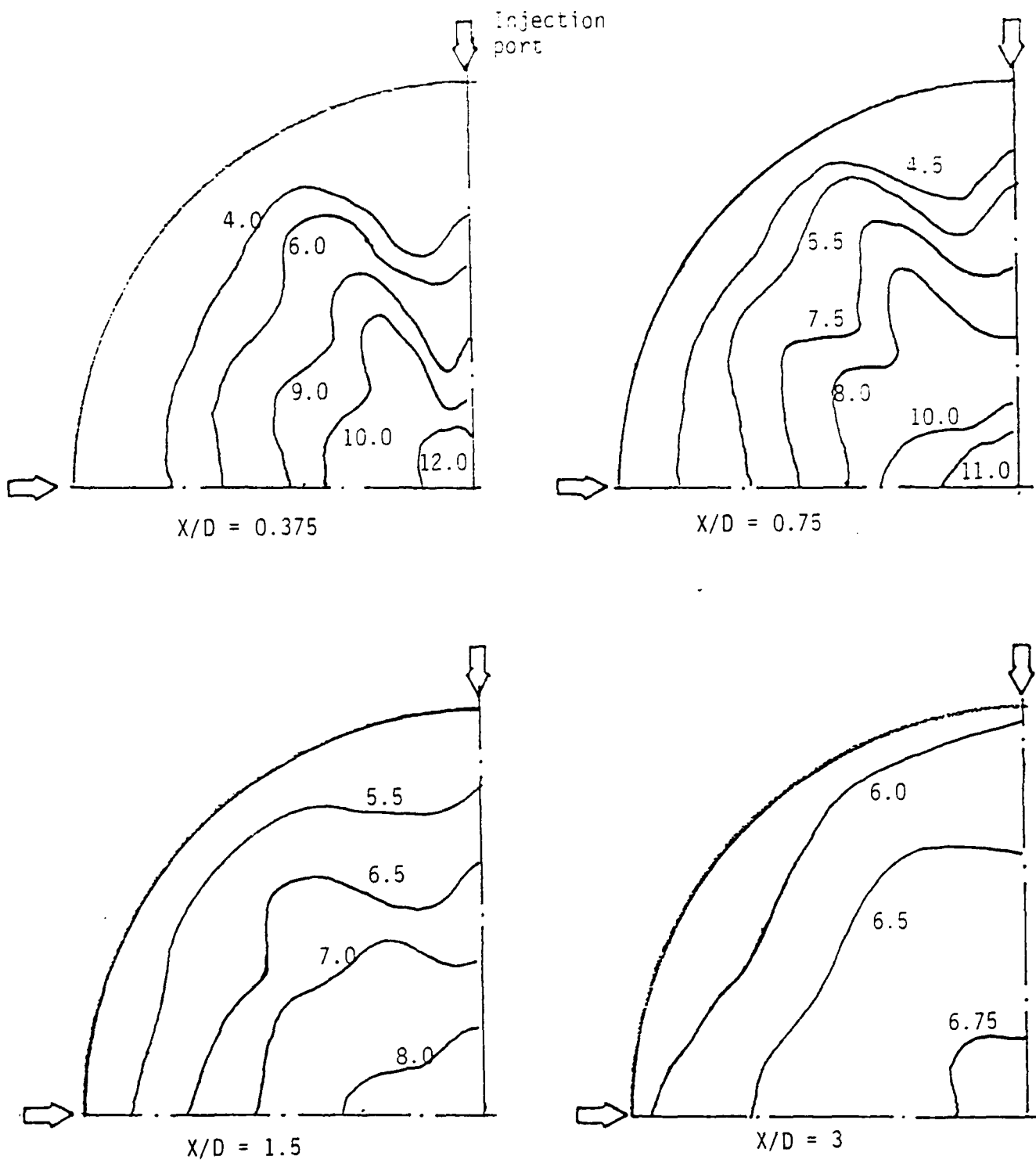


Fig. 8 Contour maps ($\theta\%$ = constant) for 4-jet air mixer
 ($V_j/V_o = 5$, $d = 0.5''$)

deviation are markedly different. Such difference raises the question about the validity of simulating heat transfer process with mass transfer process in water mixers. Such a point will be discussed in greater detail in Section VI.

C. The Diameter Ratio d/D

The effect of diameter ratio was examined in the gas mixer only and the purpose was to compare the measured data with previous investigation. The experiment was thus confined to a single jet being issued perpendicularly to a cross flow and the jet diameter was varied only once.

The measured temperature profile at three locations on the symmetrical plane of the jet is shown in Fig. 9 for two different diameter ratios. As can be seen from Fig. 9, the jet with larger diameter would penetrate deeper into the primary flow and the temperature distribution is more uneven. This observation is in agreement with that given by Kamotani et al. (1972).

The jet diameter would thus affect mixing in two opposite ways. An increase in jet diameter would increase the size of one constituent and thus the unevenness. On the other hand, it will also promote penetration and turbulence in the flow field which is beneficial to mixing. To achieve the best result of mixing, a compromise between these two effects will have to be made. According to the recommendation of Simpson (1975), the desired diameter ratio for a multi-jet cross flow mixer should be roughly equal to $1/16$. For a single-jet pipe mixing, Forney (1986) suggested that at a given jet velocity the diameter of the jet

$d/D = 1/16$
 $d/D = 1/12$

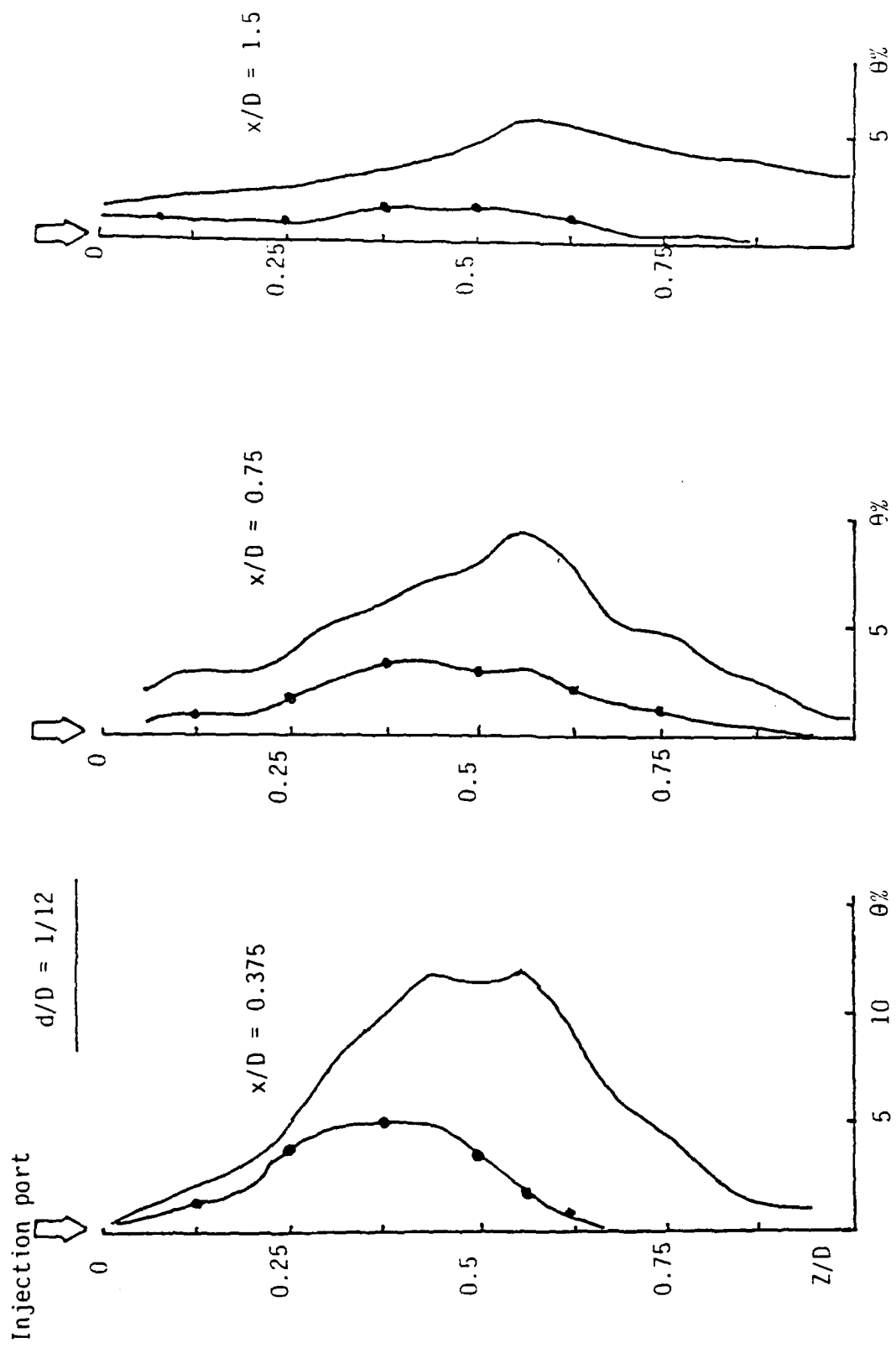


Fig. 9 Temperature profile on the plane of symmetry for a single jet injection

should ensure a penetration to the center line of the pipe. Since this work is dealing with multi-jet mixer, the recommendation from Simpson (1975) is subsequently adopted.

D. The Velocity and Density Ratio

The effect of these two ratios was examined rather thoroughly in the works of Holdeman et al. (1977) and Kamotani et al. (1972, 1974). However, these works were confined to multi-jet mixing in a rectangular duct. To complement the research work mentioned above, experiments were carried out in this work to obtain results of mixing in circular ducts. This type of results, to the author's knowledge, has not been published previously.

The construction of the mixers used in this study does not allow for significant variation of density, as can be seen from Tables 1 to 3. In fact, since the change in density is so small, all the tests carried out in either the gas mixer or water mixer can be regarded approximately as constant. Thus the effect of density ratio will be assessed by experimental evidence provided by other investigators.

The works of Kamotani et al. (1972, 1974) and Holdeman et al. (1977) indicated that the density and velocity were acted together in a combined fashion through a momentum flux ratio $J (= \rho_j V_j^2 / \rho_o V_o^2)$. Apart from contributing to the momentum flux ratio, they conclude that the effect of density ratio alone was secondary when the density ratio was within their experimental limit of 2.8. The work of Vranos et al. (1988) also tended to support the above contention in which the jet trajectory was found to follow the following

equation

$$\frac{Z}{d} = 0.76 J^{0.52} \left(\frac{\rho_j}{\rho_o} \right)^{0.11} \left(\frac{X}{d} \right)^{0.29} \quad (10)$$

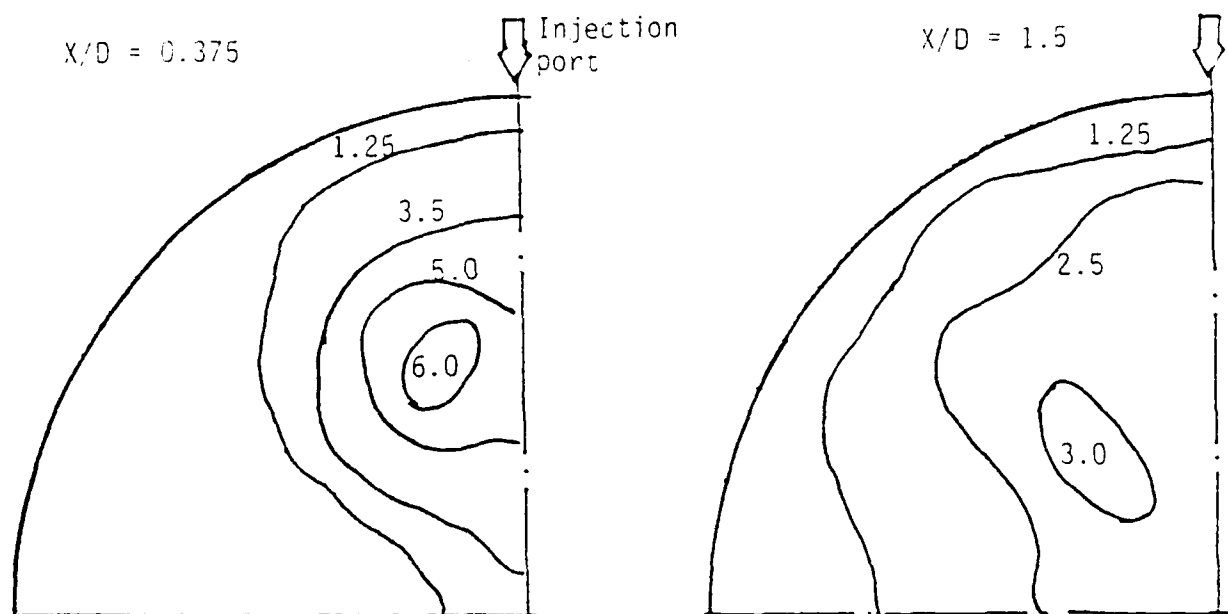
According to Eq. (10), a twenty-fold increase in density ratio will bring about less than 40% change in jet penetration. However, the effect of density ratio is still uncertain if the ratio exceeds the experimental limit mentioned above and further work in this regard is recommended.

The experimental results quoted in the previous paragraph are certainly admissible to the result of Eq. (6). The experimental results just show that the density ratio and the velocity ratio in Eq. (6) should act in a combined fashion to form a momentum flux ratio when the density ratio is not larger than 2.8.

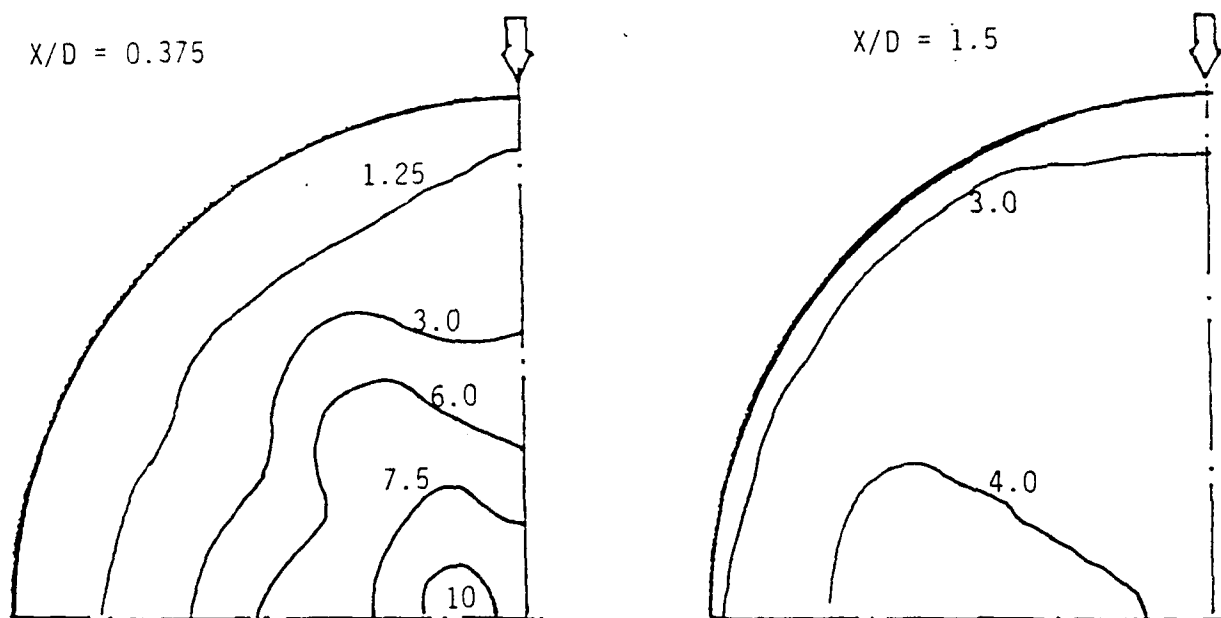
The equipment used in this work did not allow for significant variation in density, and within the experimental limit no dependence on density ratio was detected. The perception that the two ratios should be combined was thus adopted in the present study. The contour maps obtained from two different momentum flux ratios are shown in Fig. 10.

As can be seen in Fig. 10, the jets with higher J impinge with each other at an earlier stage in comparison with the jets with lower J . The standard deviations of average concentration σ at $x/d = 50$ are 21% (for $J = 49$) and 45 (for $J = 14$). Thus an increase in velocity ratio is beneficial to the degree of mixing. Whether the mixedness will increase indefinitely with J remains to be verified.

The positive effect of J on mixing is probably true for multi-



(a) Momentum flux ratio $J = 14$



(b) Momentum flux ratio $J = 49$

Fig. 10 Contour maps at two different momentum ratio for two-jet mixing ($d = 0.5''$)

jet mixers only. In this case the jets will collide with one another at the central axis of the mixer, and the turbulent intensity downstream of the point of coalescence increases with the momentum of the jets. An increase in turbulence level will certainly be beneficial to the mixing process. On the other hand, if the mixer has only one jet being issued into the primary flow, then an increase in momentum flux ratio may not be beneficial to the mixing result. (Forney, 1986).

E. Effect of Reynolds Number and Richardson Number.

As can be seen from Tables 1 to 3, the Reynolds number ranges between 12,000 to 30,000 in the two water mixers, and the variation covers only the test with single jet injection. On the other hand, experiment on the gas mixer is more complete and the Reynolds number ranges between 29,500 and 88,000 for the one jet, two jets and four jets injection. The results being presented in this report will thus be focused on the air mixer and one typical result is shown in Fig. 11.

For the gas mixer, it was found that when the Reynolds numbers were within the experimental limits, no consistent correlation between the standard deviation of the scalar field and Reynolds number was detected for the three injection arrangements. This observation seems to be in agreement with those of Rathgeber et al. (1983) and Kamotani et al. (1972). The observation made by those investigators, however, was at a distance fairly close to the injector port where the identity of the jet was not lost. In the present study, however, most of the observations were made in the

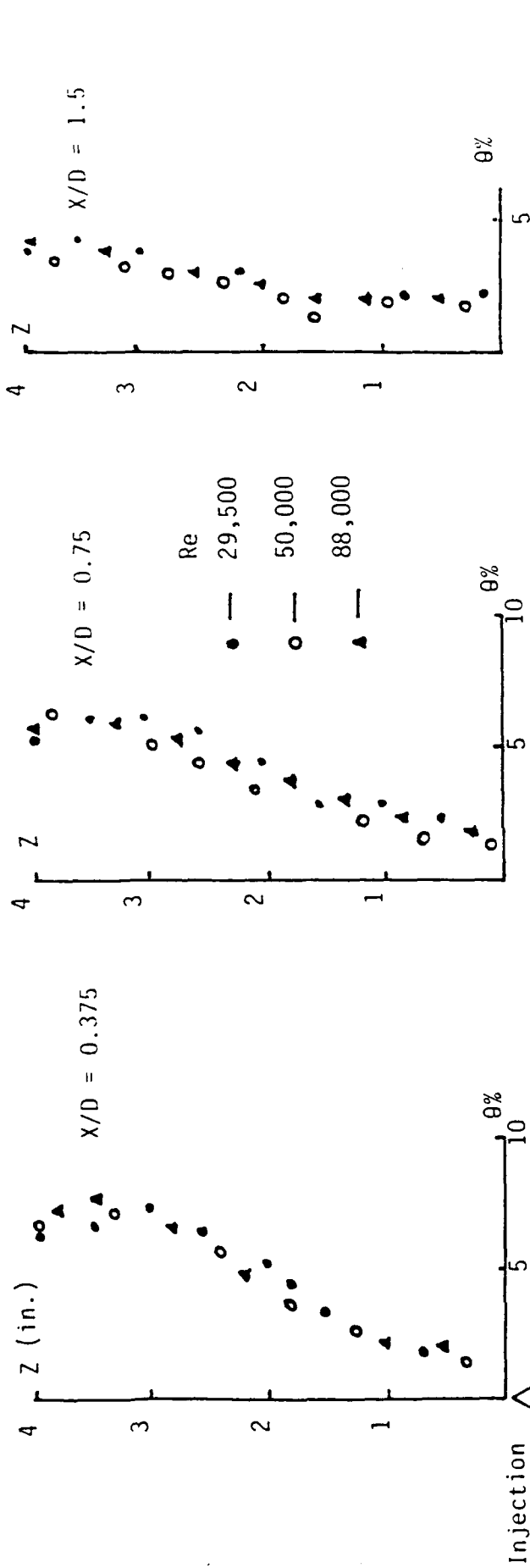


Fig 11(a). Temperature profile at plane of symmetry for two-jet mixing

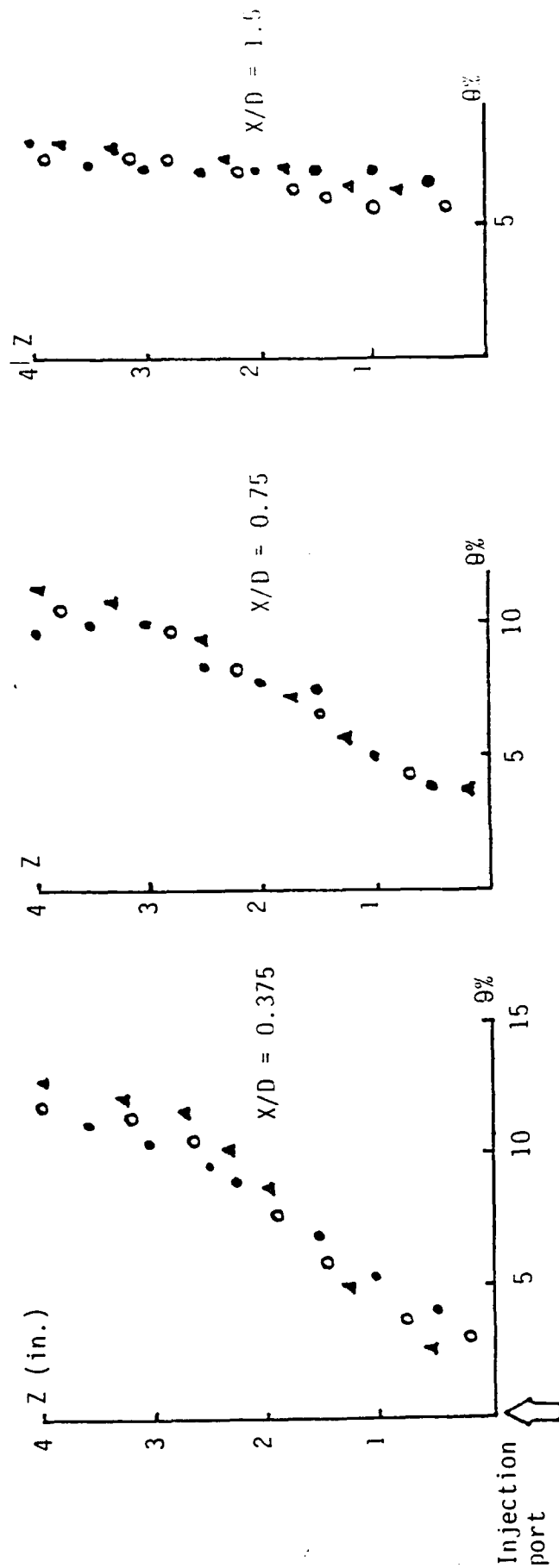


Fig 11(b). Temperature profile at plane of symmetry for four-jet mixing

far field after the merging of the jets.

An increase in Reynolds number in the gas mixer would bring about two opposite effects. The first is to increase the turbulence level in the flow (which will enhance the mixing process); the second is to reduce the resident time of the fluid in the mixer, which tends to reduce mixedness. The insensitivity of the scalar field on Reynolds numbers indicates that the two opposite effects mentioned above are canceling each other. This contention is in agreement with the theoretical analysis obtained by Beek et al. (1959). In Beek's analysis, the mixing was measured by the intensity of segregation I_s , a parameter referring to fluctuation with time at a spatial point. While in the present work, the mixing is indicated by the deviation of the average value with respect to the cross-sectional average.

For the water mixers, the experimental data can afford only a comparison for a single jet with Reynolds number ranging from 12,000 to 30,000. Such a result is shown in Fig. 12 which shows a positive effect of increasing Reynolds number on mixing. It must be pointed out that the number of experiments being carried out is not large enough and thus the result cannot be regarded as a solid proof that the degree of mixing in a liquid mixer is sensitive to Reynolds numbers, as predicted by Beek et al. (1959). According to that analysis the increase in Reynolds number would reduce the time constant t_c at a faster rate than the decrease in resident time t_r , thus the degree of mixing would improve with increase in Reynolds number.

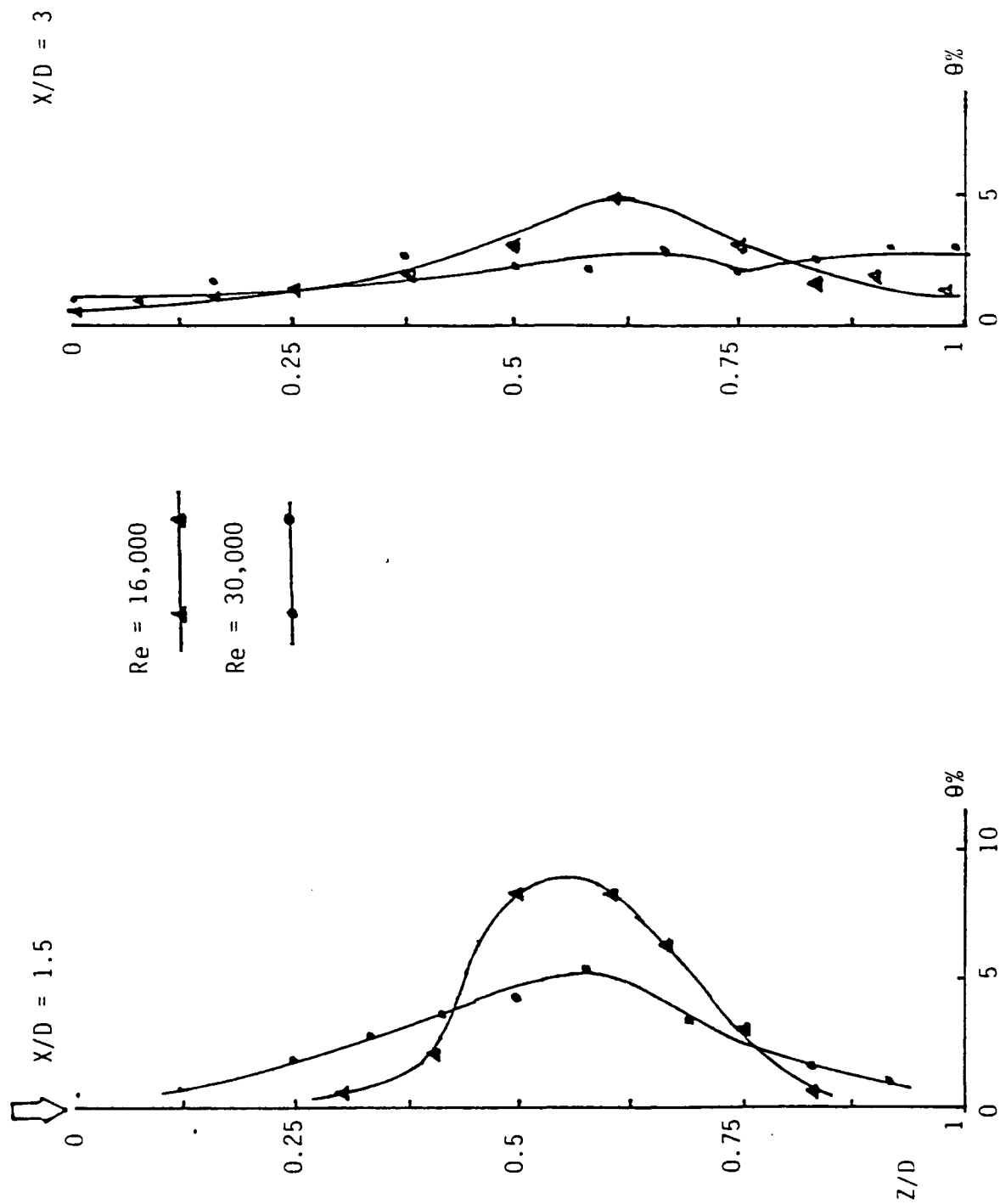


Fig. 12 Concentration profile in the plane of symmetry for a single jet water mixer ($V_j/V_0 = 4$)

The effect of Richardson number (an indicator of buoyancy effect), was examined with the data from the two water mixers by changing the concentration of the salt brine and the size of the mixer. Since the attainable density difference between the primary and the injected fluid is relatively small with the present experimental technique, the magnitude of the Richardson number covered in this work is also relatively small. With the density ratio varied from 1.02 to 1.15 at a Reynolds number of 16,000, no consistent correlation with Richardson number was found in the present work (Fig. 13). According to the work of Forney et al. (1979), the buoyancy effect would not be significant if the densimetric Froude number is larger than the jet-to-primary fluid velocity ratio, i.e.

$$\frac{V_j}{\sqrt{gd \frac{\rho_0 - \rho_j}{\rho_0}}} > \frac{V_j}{V_0}$$

or

$$V_0 > \sqrt{gd \frac{\rho_0 - \rho_j}{\rho_0}} \quad (11)$$

Since the condition specified in Eq. (11) is always satisfied in this work, thus the buoyancy effect is probably not significant enough to be detected by the present experimental techniques.

F. The Effect of Stanton Number

The Stanton number St is related to the ratio of convection heat transfer rate across the jet surface to the heat carried by the jet stream. In a jet mixer, the Stanton number is derived from the energy equation in the near field. As is well known in heat transfer texts, the Stanton number ($h/(\rho C_p V)$) is a function of

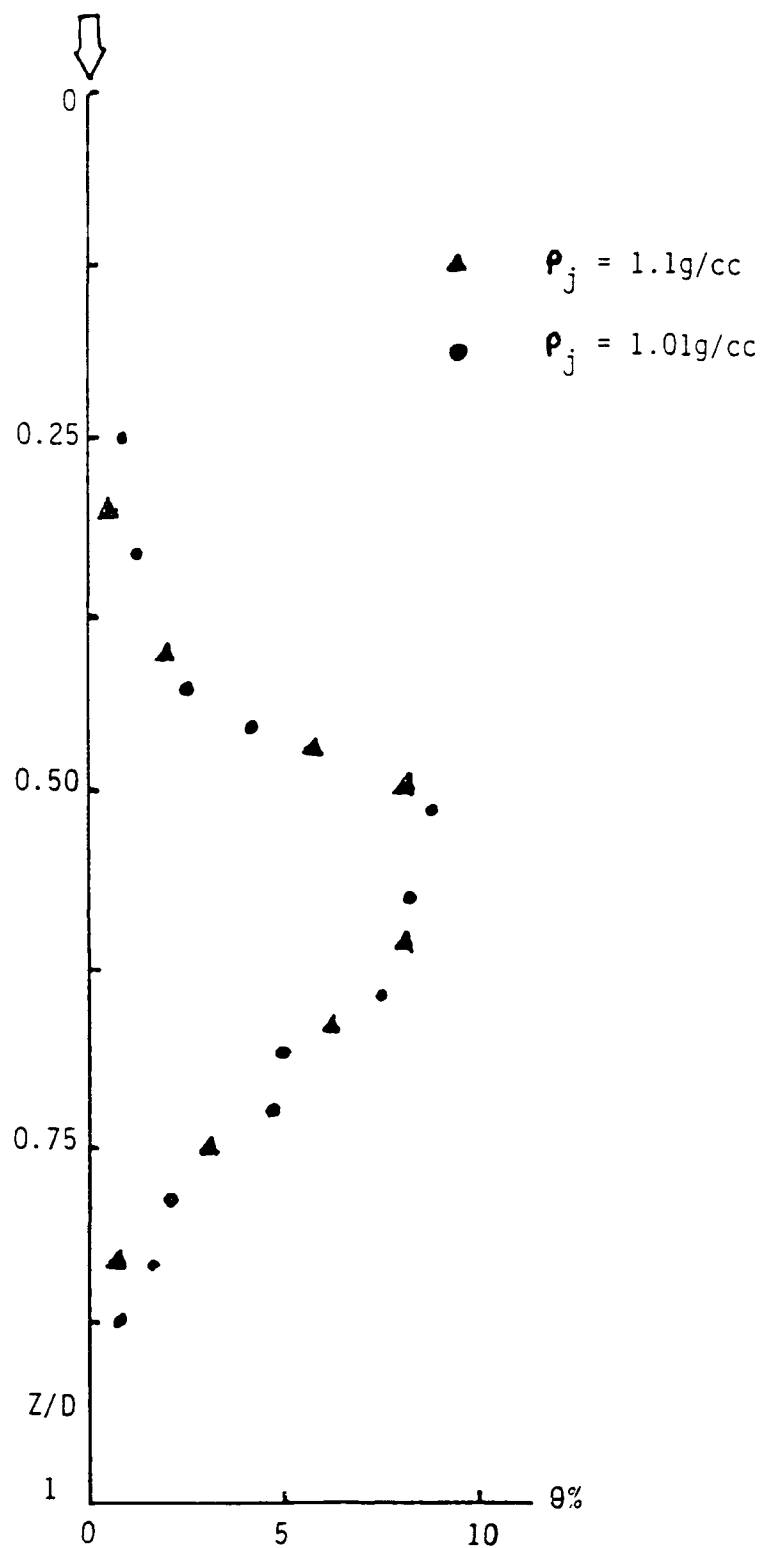


Fig. 13 Concentration profile at the plane of symmetry for single jet water mixer ($Re = 16000$, $X/D = 1.5$)

Reynolds number and Prandtl number. In gas mixers, the scalar field is insensitive to Reynolds number, thus the effect of Stanton number is represented by the Prandtl number alone. In the case of water mixer, the Reynolds number seems to have some bearings on the far field, thus to match the Stanton number requires the matching of both the Reynolds number and Prandtl number. Since the mixing result in a gas mixer is insensitive to Reynolds number, one can thus find the effect of Prandtl number by comparing the near-field scalar quantity in the gas mixer and the water mixer, even though the Reynolds number in the former is not equal to that in the latter.

The mixing in the far field depends on two factors, the first being the initial degree of segregation (represented by the scalar quantity at the end of the near field) and the second being the time ratio t_r/t_c . The time constant of mixing t_c for a gas mixer is a function of Pr , L_s and Re (see Eq. 3) and a change in Re would also change t_r . The time ratio, however, is insensitive to Reynolds number since t_r and t_c change with Reynolds number at approximately the same manner. For the mass transfer process in the water mixers, the time constant t_c is a function of Sc , L_s , Re and ν in a manner depicted in Eq. (4). With Eq. (4), the time t_c and t_r change differently with Re and thus the time ratio is found to be dependent on Reynolds number. Moreover, the time constant for the mass transfer process in the water mixer is much larger than that of the gas mixer because the Sc for salt brine is 745 (Kays, 1966) and the Prandtl number for air is 0.7 only. Thus by

comparing in scalar measurements in the far field of the gas and water mixers, one can assess the effect of Pr and Sc on mixing.

The measured scalar field for the gas mixer and the water mixers are plotted in Fig 14 (for one jet), Fig. 15 (for two jets), and Fig. 16 (for four jets). The scalar quantity in the water mixers is much non-uniform in both the near field and the far field. In the near field, the difference is caused by the difference in Prandtl number (with Pr of water equal to 7 and Pr for air equal to 0.7). In the far field, the difference is caused by the difference in t_c . For mass transfer in water, the time constant is much larger than the time constant for the heat transfer process in the gas mixer (the Sc is more than 1000 times larger than Pr). This implies the mixing in water is much slower than that in air. This conclusion is in agreement with that given by Simpson (1972).

VI. VALIDITY OF SIMULATION WITH WATER TUNNEL.

Experiment evidence from the present work and previous investigations indicate that there are several parameters pertinent to the results of mixing. For the near field, the parameters are 1) diameter ratio d/D , 2) number of injection ports, 3) momentum ratio J , and 4) the Prandtl number. For the far field, the mixing is determined by 1) the initial scale of segregation, 2) the dissipation rate, 3) the Prandtl number (for heat transfer) or Schmidt number (for mass transfer), and 4) the Reynolds number (for water mixer). The effect of the second group of parameters can be summed up by a time ratio t_r/t_c . These parameters can be

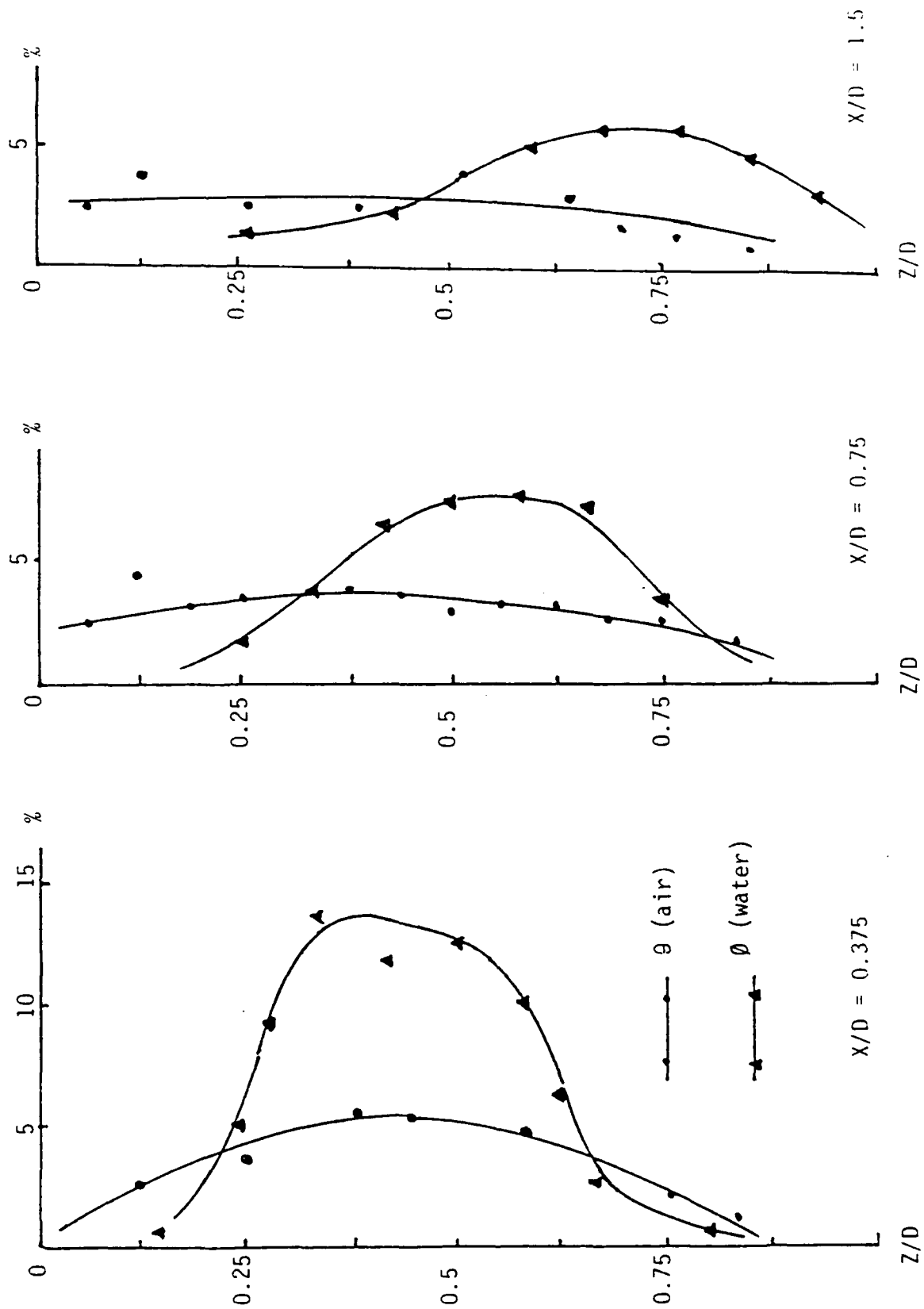


Fig. 14 Comparison of scalar field at plane of symmetry for single-jet mixer
($V_j/V_o = 5$, $Re = 30,000$)

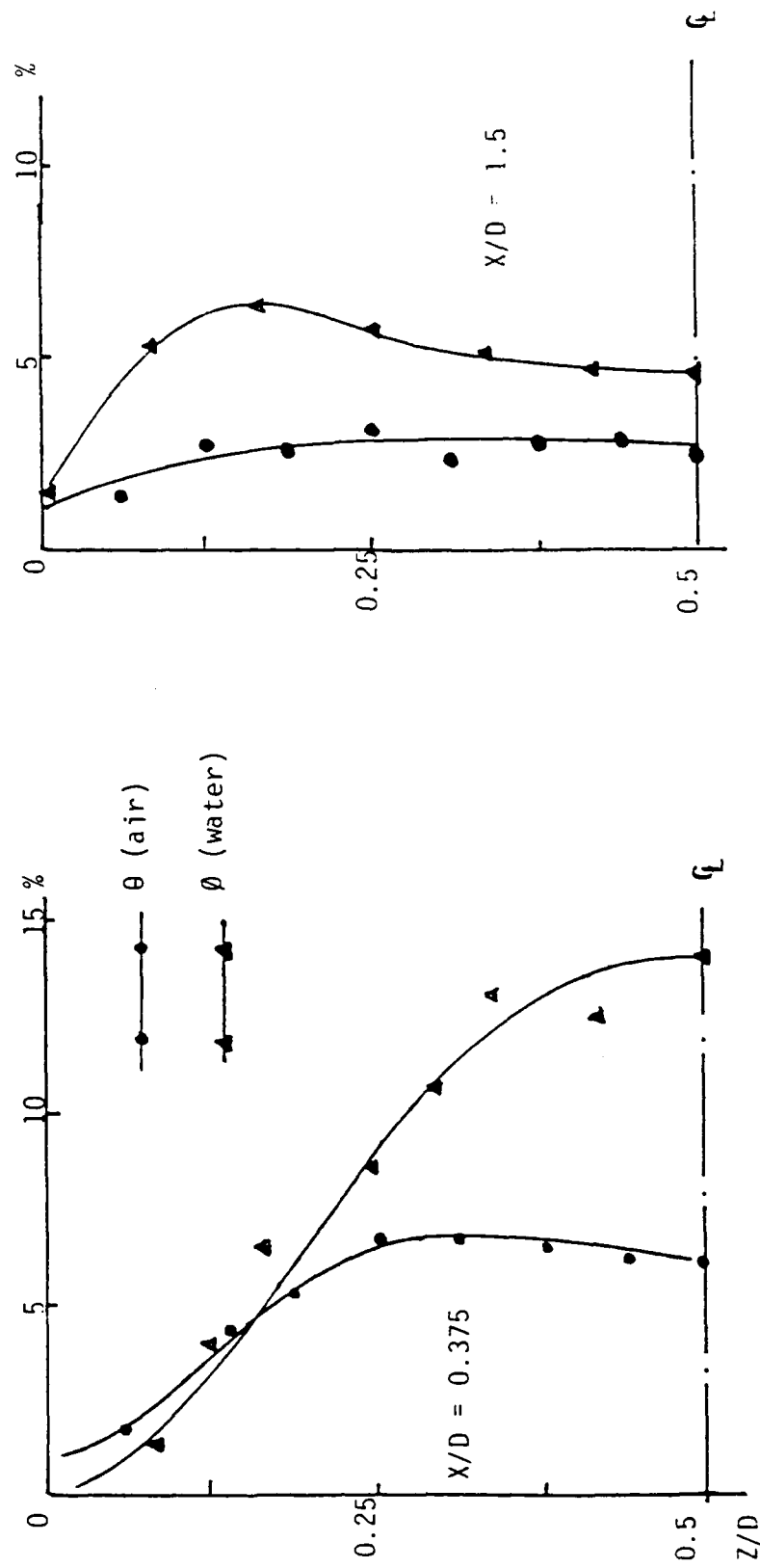


Fig. 15 Comparison of scalar field at plane of symmetry for two-jet mixer
 ($V_j/V_o = 5$, $Re = 30,000$)

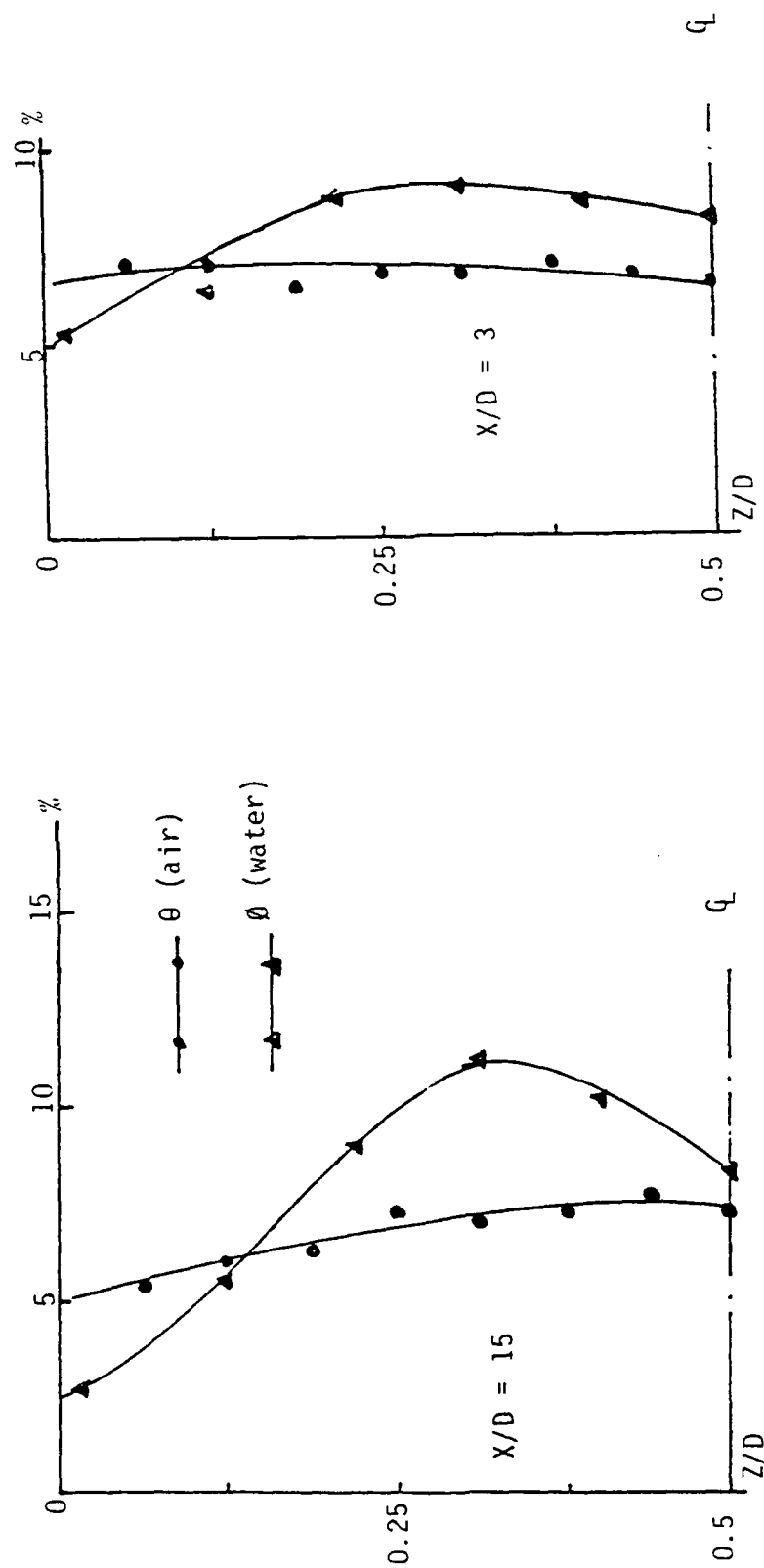


Fig. 16 Comparison of scalar field at plane of symmetry for four-jet mixer
($V_j/V_o = 5$, $Re = 16000$)

used as the similarity criteria to examine whether mass transfer process in a water mixer can be used to simulate the heat transfer process in an air mixer.

As far as the near field is concerned, the four dimensionless parameters can easily be matched between the two types of mixers except the Prandtl number. Such effect can be seen in Figs 13, 14 and 15 when $x/d = 6$. The scalar field is much less uniform for water mixer because the mass diffusion coefficient is much smaller than the thermal diffusivity. These results show that water mixer cannot simulate the air mixer in the near field.

In the far field, the similarity parameter is the time ratio t_r/t_c . Again because of the difference between Pr and Sc , the time constant for mass transfer in a water mixer is much larger than that for heat transfer in an air mixer. In other words, the speed of turbulent mixing in water is much slower than that in gas. It may first appear that similarity is attainable by adjusting the length of the mixer to achieve identical t_r/t_c . Such a method would be a valid one if t_r/t_c for the two types of mixer would remain the same when the operating conditions change. However, the analytical result of Beek et al. (1959) showed that this was not the case.

According to Beek et al., the mixing distance (a distance required to attain certain degree of mixing) in an air mixer is quite sensitive to the number of injection ports, but is rather insensitive to the Reynolds number. On the other hand, the result of mixing in a water mixing is quite different. It is rather insensitive to the number of injection ports, but very sensitive to

the Reynolds number. Such a difference is caused by the difference in the correlation between t_c and other flow parameters (see Eqs. (3) and (4)) for the two fluids. Thus for a given Reynolds number and injection ports, one may match the time ratio of the two mixers to obtain similarity in the far field. However, similarity would break down in the same mixers if the flow rate or the number of injection port is changed.

Thus the validity of using mass transfer process in a water mixer to simulate the heat transfer process in an air mixer is open to question. Indeed, our experiment measurement has shown that the scalar fields from these two mixers are quite different.

VII. CONCLUSIONS

The conclusions from the studies are:

1. Parameters pertinent to the mixing in the near field include:
 - a. the diameter ratio d/D and the number of injection ports
 - b. the momentum flux ratio $\rho_j V_j^2 / \rho_o V_o^2$
 - c. the Prandtl number (for heat transfer) and Schmidt number (for mass transfer). These two parameters were found to play a significant role in the spread of the scalar fields
2. The following parameters are found to be relatively unimportant to the mixing process
 - a. The Reynolds number: within the range of the experiment, i.e., $30000 < Re < 88000$ for gas mixer and $12000 < Re < 30000$ for water mixer, the effect of Reynolds number on the near field is insignificant.
 - b. The Richardson number. When the velocity of the primary

flow V_0 satisfies the following condition:

$$V_0 > \sqrt{gd \frac{\rho_0 - \rho_i}{\rho_0}}$$

the effect of Richardson number is found to be negligible.

3. The mixing study in the far field is in general agreement with the prediction of Beek et al. (1959), i.e.
 - a. the mixing in gas is sensitive to the number of injection ports, but is rather insensitive to Reynolds number.
 - b. the mixing in liquid is rather insensitive to the number of injection ports, but is sensitive to the Reynolds number.
4. The heat transfer process in a gas mixer cannot be duplicated closely by the mass transfer in a water mixer; and the difference between the values of the Prandtl number and Schmidt number is the major cause of this problem.

VIII. RECOMMENDATIONS

It is thus recommended that gas should be used as the test medium. If mass transfer is used to simulate the process, then the Schmidt number of the test mediums should be close to the Prandtl number. The present study should be further extended to include the following aspects.

1. To measure the mixing distance required to accomplish the desired degree of mixing. To do so the set-up used in this work needs to be upgraded, such as increasing the length of the mixing chamber and the temperature of the secondary fluid.

2. To measure the fluctuating scalar field. This measurement would provide a direct comparison with Eq. (5).

3. To evaluate the effect of density ratio. The effect of density is known to be insignificant when the density ratio is less than 2.8. However, the density ratio in a prototype mixer can reach as high as 20, and the effect of density ratio at the range is unknown and further experiment is required.

REFERENCES

- Beek, John, Jr. and R.S. Miller, Turbulent Transport in Chemical Reactors, Chemical Engineering Progress Series, Vol. 25, No. 55, 1959, pp. 23-28.
- Brodkey, R.S. The Phenomena of Fluid Motions, Addison-Wesley Publishing Co., Inc., 1967.
- Corrsin, S., Simple Theory of an Idealized Turbulent Mixer, AICHE Journal, Vol. 3, 1957, pp. 329-330.
- Corrsin, S., The Isotropic Turbulent Mixer: Part II, AICHE Journal, Vol. 10, 1964, pp. 870-877.
- Forney, L.J., and T.C. Kwon, Efficient Single-Jet Mixing in Turbulent Tube Flow, AICHE Journal, Vol. 28, 1979, pp. 623-690.
- Forney, L.J., and H.C. Lee, Optimum Dimensions for Pipeline Mixing at a T-Junction, AICHE Journal, Vol. 28, No. 6, 1982, pp. 980-987.
- Forney, L.J., Jet Injection For Optimum Pipeline Mixing, (Chapter 25, Encyclopedia of Fluid Mechanics, Vol. 2, N.P. Chermisinoff Ed.,) Gulf Publishing Co., 1986.
- Holdeman, J.D., and R.E. Walker, Mixing of a Row of Jets with a Confined Crossflow, AIAA Journal, Vol. 15, 1977, pp. 243-249.
- Kamotani, Y. and I. Greber, Experiments on a Turbulent Jet in a Cross Flow, AIAA Journal, Vol. 10, 1972, pp. 1425-1429.
- Kamotani, Y. and I. Greber, Experiments on a Confined Turbulent Jets in Cross Flow, NASA CR-2392, 1974.

- Kays, W.M., Convective Heat and Mass Transfer, McGraw-Hill, Inc.,
1st Ed., 1966.
- Kline, S.J., Similitude and Approximation Theory, Springer-Verlay,
Inc., 1986.
- Lee, L.W., The Design of Jet Mixers for an Arc Heater: An
Experimental Approach, 1989 USAF-UES Summer Faculty Research
Program, Final Report, 1989.
- Mckelvey, K.N. et al., Turbulent Motion, Mixing, and Kinetics in a
Chemical Reactor Configuration, AICHE Journal, Vol. 21, 1975,
pp. 1165-1176.
- Moussa, Z.M. et al., The Near Field in the Mixing of a Round Jet
with a Cross-Stream, J. Fluid Mechanics, Vol. 80, 1977, pp.
49-80.
- Rathgeber, D.E. and H.A. Becker, Mixing between a Round Jet and a
Transverse Turbulent Pipe Flow, The Canadian Journal of
Chemical Engineering, Vol. 61, April, 1983, pp. 148-157.
- Shope, F.L., Private communication, 1989.
- Simpson, L.L., Turbulence and Industrial Mixing, Chemical
Engineering Progress, Vol. 70, No. 10, 1974, pp. 77-79.
- Simpson, L.L., Turbulence in Mixing Operation, R.S. Brodkey Ed.,
(New York, Academic Press), 1975, pp. 279-330.
- Vranos, A. and D.S. Lincinsky, Planar Imaging of Jet Mixing in
Crossflow, AIAA Journal, Nov., 1988, pp. 1297-1298.
- Wright, S.J. Mean Behavior of Buoyant Jets in a Cross Flow, J.
Hydraulic Div., Trans. ASCE, HY5, 1977, pp. 499-513.

Report # 16
760-6MG-099
Prof. Arthur Mason
Report Not Publishable

RESEARCH INITIATION PROGRAM

FINAL REPORT

LASER-INDUCED FLUORESCENCE OF NITRIC OXIDE

Submitted to the Sponsor

Air Force office of Scientific Research

Conducted by the

Universal Energy System, Inc.

Contract Number: F49620-88-C-0053/SB135881-0378

Prepared by

Chun Fu Su

Department of Physics and Astronomy

Mississippi State University

Mississippi State, Mississippi

December 1990

ABSTRACT

Preliminary computational models of the laser-induced fluorescence spectrum of nitric oxide have been developed for two different electronic transitions:

$B^2\Pi (v'=7) - X^2\Pi (v''=0)$ and $D^2\Sigma (v'=0) - X^2\Pi (v''=1)$. These codes have been developed for a single-photon process with the spectrum displayed either as a stick spectrum or displayed using a Lorentzian, Gaussian, or Voigt line shape. Laser-induced fluorescence spectra of nitric oxide at room temperature and at high temperature (up to 623°K) with various pressures have been recorded. Comparison of the experimental results with predicted spectra will be presented herein.

ACKNOWLEDGEMENTS

I wish to express my appreciation and gratitude to the Air Force System command, the Air Force Office of Scientific Research, and the Arnold Engineering Development Center for sponsorship of this research. Sincere thanks are extended to the Universal Energy System for their concern and help in all aspects of this program. I greatly appreciated the collection of the observed LIF spectra by Mike Smith, the operation assistance of Billy McClure and Linwood Price, and the computing facilities at AEDC and Mississippi State University. Finally, I deeply thank Donnie Williams, Section Head, for his support and encouragement throughout this project.

I. INTRODUCTION

The NO molecule is a stable diatomic molecular radical. The molecule is a key molecule in the chemistry of the upper atmosphere and an important but dangerous pollutant. It has also been detected in an interstellar molecular cloud. Moreover, this molecule is important to combustion chemistry and turbulent flows. Accurate measurements of temperature and species concentration provide the fundamental diagnostic input to characterize combustion system and turbulent flows. It is obvious that this molecule can provide many interesting insights to various fundamental processes.

The development and application of laser-induced fluorescence as a diagnostic technique has been underway for approximately twenty years. This experimental method can be used to make non-intrusive measurements in combustion systems and arc-heated flow fields. Measurements inside a combustion system or an arc-heated flow can be used to determine the species density and temperature as well as to assign the molecular transitions.

Study of laser-induced fluorescence of nitric oxide was therefore decided on as the subject of this research project. Single photon excitation rather than multiphoton excitation was used. The ArF excimer laser system, Lambda Physik EMG 150 MSC, installed at AEDC was used for this project.

II. OBJECTS OF THE RESEARCH EFFORT

For the overall research project, several individual tasks were done. In particular, they were:

1. Development of computational models for electronic transitions;
2. Observation of rotationally resolved laser excitation spectra at different temperatures and at various pressures;
3. Observation of vibrationally resolved laser excitation spectra at different temperatures and at various pressures.

III. ENERGY LEVELS

In order to understand the behavior of the fluorescence signals that can be affected by temperature, pressure, and spectroscopic constants, computational models were needed. It was decided to observe and characterize the fluorescence spectra at room temperature and at high temperature (1200°K).

According to the relative intensity calculations, it was found that the rotational transition of the $B^2\Pi(v'=7) - X^2\Pi(v''=0)$ electronic transition would be important at room temperature, those of the $D^2\Sigma(v'=0) - X^2\Pi(v''=1)$ electronic transition would be dominant at high temperature, and those of these two electronic transitions would appear together at moderately high temperature. Therefore, three different computational models of fluorescence spectra were needed for this project. The important equations for the rotational energy levels, rotational transition strengths, and line shapes will be briefly described below.

The total energy of a diatomic molecule, excluding translation, is

$$E_t = E_e + E_v + E_r,$$

where E_e , E_v , and E_r respectively stand for the energies of the electronic state, vibrational state, and rotational state. The second two terms on the right are well known. They are

$$E_v = W_e(v + \frac{1}{2}) - W_e X_e(v + \frac{1}{2})^2 + W_e Y_e(v + \frac{1}{2})^3, \quad (1)$$

and

$$E_r = K(K+1)B_v - K^2(K+1)^2 D_v,$$

where v and K are vibrational and rotational quantum numbers, W_e is the vibrational energy expressed in cm^{-1} . $W_e X_e$ and $W_e Y_e$ are the anharmonic constants, B_v and D_v are, respectively, the rotational constant and centrifugal distortion constant of the vibrational state v .

For the NO molecule the electronic spin has the value $\frac{1}{2}$ which produces doublets. One labels the two components of the doublet by subscripts 1 and 2, so that

$$F_1(K); f_1(K):J = K + \frac{1}{2},$$

and

$$F_2(K); f_2(K):J = K - \frac{1}{2},$$

where K and J , respectively, represent the pure rotational quantum number and the coupled quantum number. $F_1(K)$ and $F_2(K)$ are used to represent the rotational levels of the $^2\Sigma$ electronic state, while $f_1(K)$ and $f_2(K)$ for those of the $^2\Pi$ electronic state.

Besides the spin effect, the rotational energy levels will also be affected by the coupling of the spin and the momentum

created by the rotation of the nuclei. The rotational energy levels will thus be modified. They are:

$$\begin{aligned} F_1(K) &= K(K+1)B_V - K^2(K+1)^2D_V + R(K+\frac{1}{2}), \\ F_2(K) &= K(K+1)B_V - K^2(K+1)^2D_V - R(K+\frac{1}{2}), \end{aligned} \quad (2)$$

for the $^2\Sigma$ electronic state [1];

and

$$f_1(K) = [K(K+1)^{2-\frac{1}{2}} \sqrt{4(K+1)^2 + a(a-4)}]B_V - K^2(K+1)^2D_V, \quad (3)$$

$$f_2(K) = [K^{2-1+\frac{1}{2}} \sqrt{4(K^2 + a(a-4))}]B_V - K^2(K+1)^2D_V,$$

for the $^2\Pi$ electronic state [2], where R and a are the coupling constants. They are determined from the experimental spectra.

For rotational transition probabilities, two sets of equations were used. The formulas expressed explicitly in terms of the quantum numbers by Earls [3] originally obtained by Hill and Van Vleck [4] were used for the $^2\Sigma - ^2\Pi$ electronic transitions. They are:

$$\begin{aligned} \frac{P_2}{S_{21}} &= \frac{2J+1 * ((2J+1) \pm U((2J+1)^2 - 2a))}{2J+2}, \\ \frac{R_1}{Q_{12}} &= \frac{2J+1 * ((2J+1) \pm U((2J+1)^2 + 2(a-4)))}{2J+2} \cdot \frac{1}{J}, \\ \frac{Q_2}{R_{21}} &= \frac{2J+1 * [(2J+1)^2 - 2 \pm U((2J+1)^3 - 8J - 2a)]}{2J+2} \cdot \frac{1}{J}, \\ \frac{Q_1}{P_{12}} &= \frac{2J+1 * [(2J+1)^2 - 2 \pm U((2J+1)^3 - 8J + 2(a-4))]}{2J+2} \cdot \frac{1}{J}, \end{aligned} \quad (4)$$

$$\frac{P_1}{Q_{12}} = \frac{2J+1 * ((2J+1) U((2J+1)^2 - 2a))}{2J},$$

$$\frac{P_2}{Q_{21}} = \frac{2J+1 * ((2J+1) U((2J+1)^2 + 2(a-4)))}{2J},$$

$$\text{where } U^{-1} = \sqrt{(2J+1)^2 + a(a-4)}$$

and a is a coupling constant. The $+$ sign is used for the main branches such as R_2 , R_1 , etc.; while the $-$ sign is used for the satellite branches, such as S_{21} , Q_{12} , etc..

The Hond-London formulas for $\Delta\Lambda = 0$ transitions listed in the "Spectra of Diatomic Molecules" [5] were used for the $^2\Pi - ^2\Pi$ electronic transitions. They are:

$$\begin{aligned} R &= \frac{(J+1+\Lambda) * (J+1+\Lambda)}{J+1}, \\ Q &= \frac{(2J+1) * \Lambda}{J(J+1)}, \\ P &= \frac{(J+\Lambda) * (J-\Lambda)}{J}, \end{aligned} \tag{5}$$

where J and $(\Lambda=1)$ respectively stand for the coupled rotational quantum number and electronic state. There are six rotational transitions within these two electronic states:

$$R_{11}, R_{22}, Q_{11}, Q_{22}, P_{-1}, \text{ and } P_{22}.$$

The molecular concentration distribution is an important factor in the evaluation of the transition intensities. For a given electronic state, the gas density of the rotational state K in the vibrational state is

$$N_K = \frac{N(2K+1)}{Q_R Q_V} e^{-(E_R + E_V)hc/kT}, \tag{6}$$

where E_V is the vibrational energy, E_R is the rotational energy, T is the temperature, N is the total gas density (molecules/cm³),

and Q_r and Q_v are the rotational and vibrational state partition functions defined as

$$Q_r = \sum_{K=0}^{\infty} (2K+1) e^{-B_v K(K+1)hc/kT},$$

and

$$Q_v = \sum_{v=0}^{\infty} e^{-E_v(v)hc/kT}.$$

Here the term E_v stands for the vibrational energy of the v state. The transition intensity factor P_i is given by the product of Eqs. (4) or (5) and (6).

The Voigt line shape applies when both Doppler and pressure broadenings are significant, for example, when the experiment is carried out at high temperature with moderate pressure. Typical combustion applications are of this nature, therefore the Voigt line shape function will be used in this work. Several important expressions of this function will be briefly described. The transition coefficient at a wavenumber due to a transition of wavenumber λ is given as

$$I_i(\lambda) = I_i \frac{\gamma_i}{\pi} \int_{-\infty}^{\infty} \frac{\exp(-t^2)}{(x_i - t)^2 + \gamma_i^2} dt,$$

where

$$I_i = \frac{P_i}{\delta\lambda_i} \sqrt{\frac{\ln 2}{\pi}}$$

with P_i the transition intensity factor mentioned previously and $\delta\lambda_i$ is the Doppler half-width given by

$$\delta\lambda_i = \lambda_i \sqrt{\frac{2kT \ln 2}{Mc^2}},$$

$$\gamma_i = \frac{\gamma_i}{\delta\lambda_i} \sqrt{\ln 2},$$

and γ_i is the Lorentz half-width determined from experiment;

$$x_i = \frac{\lambda - \lambda_i}{\delta \lambda_i} \sqrt{\ln 2} ;$$

and t is a real number. The net coefficient at λ will be

$$I(\lambda) = \sum I_i(\lambda).$$

IV. COMPUTATIONAL MODELS

Three codes (PPTTEST.FOR, DPTEST.FOR, and TEST.FOR) were developed in this project, and they were used, respectively, to predict the fluorescence signal of the $B^2\Pi(v'=7) - X^2\Pi(v''=0)$ electronic transition at room temperature, the fluorescence signal of the $D^2\Sigma(v'=0) - X^2\Pi(v''=1)$ electronic transition at high temperature, and that of the combination of these two electronic transitions at moderately high temperature. In principle, the structure and optional functions of the first two computational codes are the same. First of all, the rotational energy levels in both electronic states were calculated. Secondly, the excitation frequency and transition intensity factor of each rotational line were calculated. Besides the transition probability, the Boltzmann distribution and Franck Condon Factors [6] were considered. Because of the same lower energy level electronic state, the partition functions for rotational and vibrational states were not used here. Third, the emission transition with its transition probability were calculated for high energy levels. Each rotational energy level from the upper electronic state will provide either six (for $D^2\Sigma - X^2\Pi$ in PPTTEST.FOR) or three (for $B^2\Pi - X^2\Pi$ in DPTEST.FOR) emission transitions. This means that the number of molecules in the higher energy level will be split into three or six parts.

Forth, the emission transition wavelength (or wavenumber) may be selected in a proper range for both prediction and plot. Each code has a function to generate the information for plotting. Four line shapes were used in each code. They were stick, Lorentzian, Gaussian, and Voigt line shapes. The stick spectrum only indicates the wavelengths and intensities. The procedures for the second and third are straightforward. However, that of Voigt line shape is relatively tedious. Some important steps will be briefly explained below.

The integral term is the most important in Voigt line shape. The magnitudes of X_i , Y_i , and t play an important role in this part. The magnitude of Y_i depends upon the Lorentz half-width γ_i and Doppler half-width $\delta\lambda_i$. The former only depends upon the experimental data, while the latter depends upon the transition wavenumber and temperature. For a given temperature and the short wavenumber region, the change of $\delta\lambda_i$ is small, so is that of Y_i . The nearly constant value of Y_i can avoid a very small magnitude in the denominator. Besides the Doppler half-width, the magnitude X_i depends on the difference between the individual transition wavenumbers and the wavenumbers in the selected region. The integrand is an even function of t (real value). The magnitude is always positive, and gradually decreases from the symmetric point ($t \approx 0$). For convenience, a summation procedure was used to replace the integration. It was found that the increment in t , Δt , and the upper and lower limits could affect the calculated values of the integral. For a given set of X_i and Y_i the calculating procedure is stopped when the

multiplied magnitude of the integrand value and increment is less than $1.0E-6$. Different values of Δt were used to check the calculated value in the summation process. When Δt was less than 0.03, less than a 1 percent change in the calculated value in the integral was found, and the value of $\Delta t=0.03$ was used in this work. The upper and lower limits in the summation were found to be around ± 3.0 .

The procedure described up to this point applies only for a given wavenumber associated with a particular transition. The same procedure is then used for the rest of the transitions. The summation of the individual contribution gives the effective spectral intensity factor at that given wavenumber (data point). The effective spectral intensity must then be calculated for each wavenumber in the wavenumber region. The selected wavenumber region was from 51600 cm^{-1} to 51880 cm^{-1} in this project, and the wavenumber separation was selected a 0.06 cm^{-1} after a number of tests. This produces about five thousand data points, which can provide a very smooth plotted spectrum. The procedure was used for both codes PPTTEST.FOR AND DPTTEST.FOR. Attempt was tried to established one code for both; however, substantial differences in calculation of transition intensity factors and rotational energy levels made it difficult. On the other hand, it was easy to combine the fluorescence rotational transitions and intensity factors respectively calculated in both codes and to plot the predicted spectra. The code TEST.FOR was used for this case.

V. EXPERIMENTAL RESULTS

a. Rotationally resolved laser excitation spectrum at room temperature

In order to understand the effect of pressure on the spectrum, various pressures were used. The laser was scanned from 192.8nm to 193.8nm to generate the electronic rotational transitions. A spectrometer set at 208nm and also a PMT with a band pass filter were used to record the rotational spectra. The spectra at pressures of 0.5 torr and 10.0 torr recorded by the PMT are shown in Figures 1 and 2, respectively. It is easy to see that the four groups of transitions were only slightly affected by pressure, and the transitions marked with (*) were significantly affected by pressure. The four groups are due to the $B^2\Pi (v'=7)-X^2\Pi (v''=0)$ transitions. The predicted spectrum with an assumed Voigt line shape is shown in Figure 3. Comparison of these three figures indicates that the observed and predicted spectra are consistent. The spectrum in the same region has been reported previously by Wodtke et al[7]. For convenience, the predicted transitions with wavelengths are listed in Table I. The line strengths of the Q-branch transitions are inversely proportional to the rotational quantum numbers; these lines, therefore, are weak in this wavelength region.

The spectroscopic constants previously determined [8] were used without any change for our models. They are listed in Table II.

b. Rotationally resolved laser excitation spectrum at high temperature

Mixtures of NO and N₂ were sent through an electrically heated tube to the cell. The NO in the mixture was approximately 5% or less. Various temperatures were reached, and the rotationally resolved laser excitation spectra at different temperatures were recorded. The recorded spectrum at the highest temperature, 623°K, is shown in Figure 4.

Comparing Figure 4 with Figures 1 and 2, indicates that the aforementioned four groups were slightly changed whereas those marked with (*) were changed more. In order to understand more about the observed spectra, the predicted spectra with a Voigt line shape were calculated for the D²Σ (v'=0)-X²Π (v''=1) transition. According to the predicted spectrum shown in Figure 5, most of the transitions marked with (*), shown in Figure 2 were due to this electronic transition. For this electronic transition, in addition to the P11, Q11, R11, P22, Q22, and R22 transitions reported by Andresen et al [9], the P12, Q12, Q21, S21, R21, and O12 transitions are also important as indicated by

Scheingraber and Vidal [10]. The assumed spectroscopic constants for these two electronic states are listed in Table II, and the predicted transitions with wavelengths are listed in Table III. The calculated spectra from both PPTEST.FOR and DPTEST.FOR are consistent with the published measurements [11]. The plotted spectrum at high temperature from DPTEST.FOR is also consistent with the observed spectrum [9].

c. Vibrationally resolved fluorescence transitions at room temperature

For fluorescence spectra, besides the rotational transitions, the vibrational transitions between two electronic states are also important. In order to understand the effects of pressure and temperature on the vibrational transitions, various pressures and temperatures were used in this research. The laser wavelength was fixed at any of the observed lines listed in Table I. The spectrometer was scanned slowly from the laser wavelength to 330nm to record the vibrational transitions. A rotational transition at the fixed input laser wavelength was excited, producing the vibrational fluorescence spectrum. The vibrational spectra at pressures of 1.0 torr, and 10.0 torr at laser wavelength set at 193.30nm are shown in Figures 6 and 7, respectively. Of the transitions listed in Table I, one or more of

four transitions R_{11} ($29\ 29\frac{1}{2} \rightarrow 30\ 30\frac{1}{2}$), P_{11} ($26\ 26\frac{1}{2} \rightarrow 25\ 25\frac{1}{2}$), R_{22} ($28\ 27\frac{1}{2} \rightarrow 29\ 28\frac{1}{2}$), and P_{22} ($25\ 24\frac{1}{2} \rightarrow 24\ 23\frac{1}{2}$) were excited. The vibrational fluorescence spectrum shown in Figure 6 is consistent with the previously reported one [12]; however, some expected vibrational transitions are still weak in Figure 6 due to effects of pressure. The high pressure vibrational spectrum shown in Figure 7 shows more vibrational transitions and stronger intensities. This is not surprising, because increasing the pressure increases the molecular density, and hence the intensities.

d. Vibrationally resolved fluorescence transitions at high temperature

The process used to heat the cell is the same as mentioned previously. The vibrational laser induced fluorescence spectra at different temperatures were recorded. The one at 623°K with laser wavelength fixed at 193.30nm is shown in Figure . It is obvious that the spectrum has even more transitions and even stronger intensities than that at room temperature. Because the rotational quantum numbers shown in Table I are moderately high, the population in these rotational energy levels should not be large at room temperature. When the temperature is increased, however, the population in these high energy levels will be increased. Moreover, the

population in the rotational energy levels of the other vibrational state ($v''=1$) will be increased at high temperature. Consequently the transitions intensities will be increased.

VI. RECOMMENDATIONS

Comparison of the experimental results recently obtained at AEDC with published reports [7,9,12] shows that the laser system works properly. Several experimental projects can be finished at AEDC in the near future using this laser system. In particular,

1. **Fluorescence spectra of rotational and vibrational transitions at higher temperatures.** So that the basic experimental results can be extended to temperature regions of interest to DoD. Those results will provide useful information for basic scientific research.
2. **Fluorescence spectra of arc-heated flows at AEDC.** Determination of temperature and molecular density is of importance for the mission of AEDC. Fluorescence spectra can be used for determination of temperature and molecular density. When finished, these experimental results will provide useful information for both basic research and diagnostic applications.
3. **Beside nitric oxide, the laser system can be used to investigate the spectra of other molecules, such as OH.**

REFERENCES

1. Van Vleck, J.H., "On σ -Type Doubling and Electron Spin in the Spectra of Diatomic Molecules", Phys. Rev. 33, 467-506 (1929).
2. Hill, E.L., and Van Vleck, J.H., "On the Quantum Mechanics of the Rotational Distortion of Multipletes in Molecular Spectra", Phys Rev. 32, 250-271 (1928).
3. Earls, L.T., "Intensities in $^2\Pi-^2\Sigma$ Transitions in Diatomic Molecules", Phys Rev. 48, 423-424 (1935).
4. See Ref. 2.
5. Herzberg, G., "Spectra of Diatomic Molecules," Van Nostrand, NY, 1950.
6. Nicholls, R.W., J. Res. Nat. Bur. Stand A. 68 535-540 (1964).
7. Wodtke, A.M., Huwel, L., Schluter, H., Meijer, G., Andresen, P., and Voges, H., "High-Sensitivity Detection of NO in a Flame Using a Tunable ArF Laser," Opt. Lett. 13, 910-912 (1988).
8. Engleman, R. Jr. and Rouse P.E., "The β and γ Bands of Nitric Oxide Observed during the Flash Photolysis of Nitrosyl Chloride," J. Mol. Spectrosc., 37, 240-251 (1971).
9. Andresen, P., Meijer, G., Schluter, H., Voges, H., Koch, A., Hentschel, W., Oppermann, W., and Rothe, E., "Fluorescence Image inside an Internal Combustion Engine Using Tunable Excimer Lasers," Appl. Opt., 29, 2392-2404 (1990).

10. Scheingraber, H. and Vidal, C.R., "Fluorescence Spectroscopy and Franck-Condon-Factor Measurements of Low-Lying NO Rydberg State," J. Opt. Soc. Am. B 2, 343-354 (1985).
11. Robie, D.C. Buck. J.P., and Bischel, W.K,. "Bandwidth and turning range of an ArFi laser measured by 1+1 resonantly enhanced multiphoton ionization of NO". Appl. opt. 29, No. 27, 3961-3965 (1990).
12. Shibuya, K. and Stuhl, F., "Single Vibronic Emission from NO $B^2\Pi$ ($v'=7$) and O₂ $B^2\Sigma_u^{-1}$ ($v'=4$) Excited by 193 nm ArF Laser," J. Chem. Phys., 76, 1184-1186 (1982).

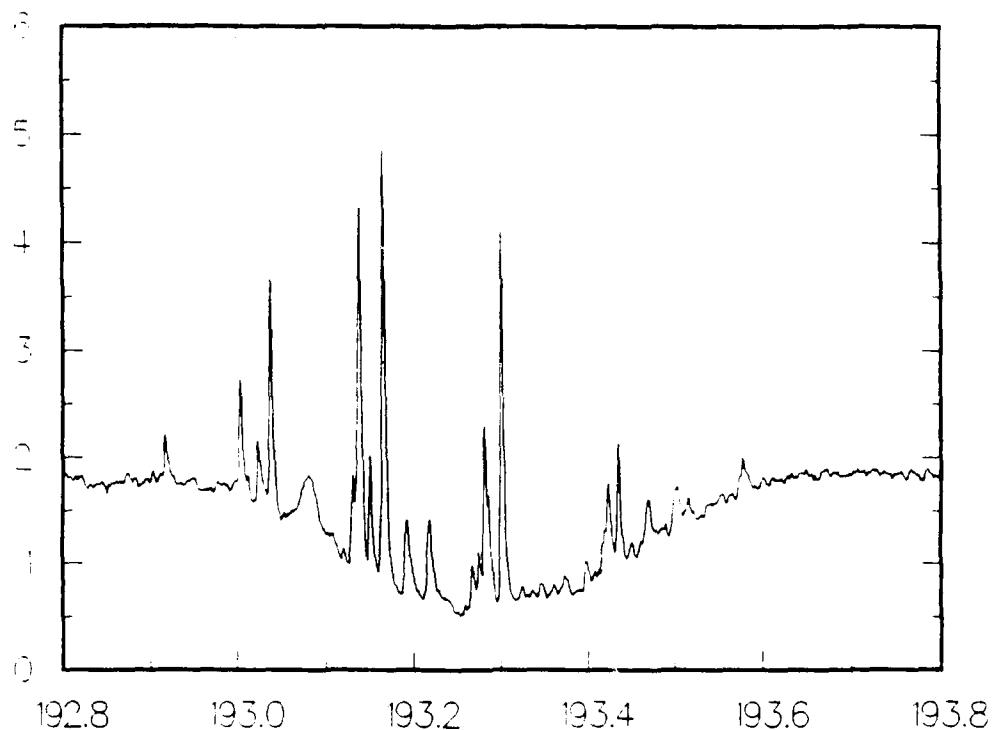


Figure 1. Observed rotationally resolved laser excitation spectrum of NO at room temperature and a pressure of 0.5 torr.

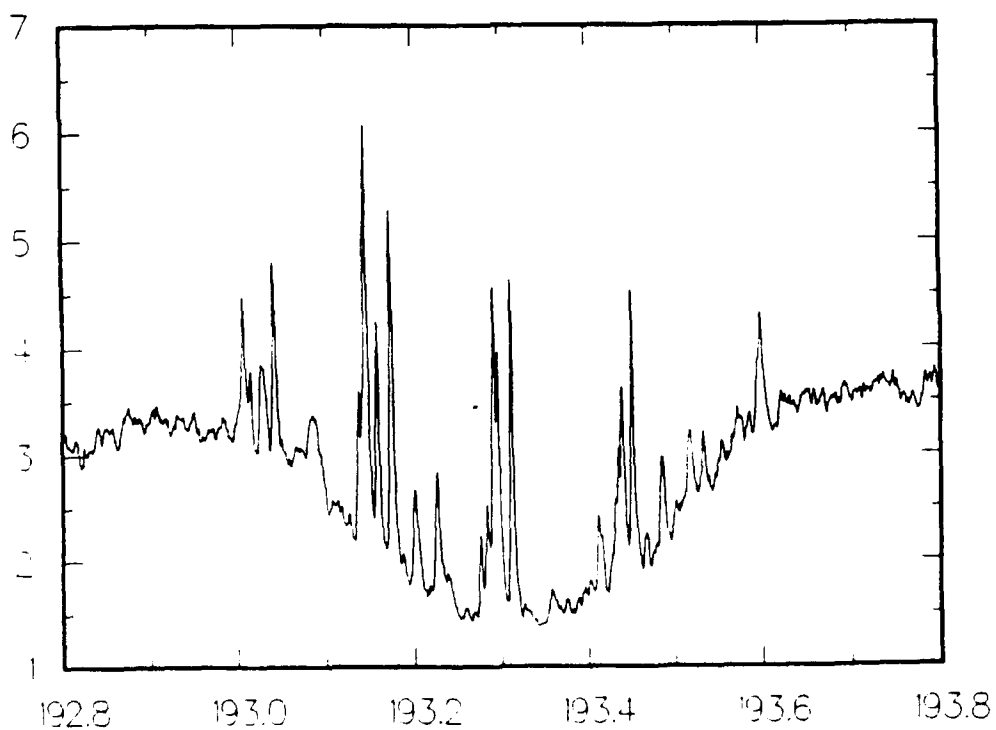


Figure 2. Observed rotationally resolved laser excitation spectrum of NO at room temperature and a pressure of 10.0 torr.

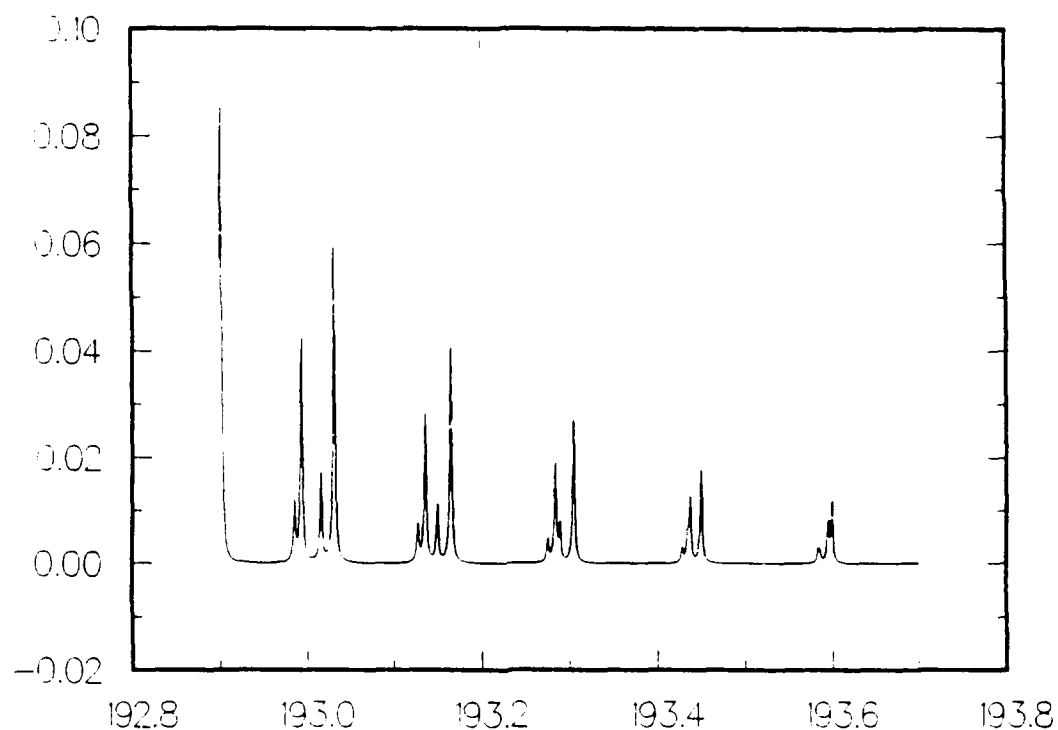


Figure 3. Predicted rotational spectrum of NO at room temperature for the $B^2\Pi(v'=7) - X^2\Pi(v''=0)$ transition.

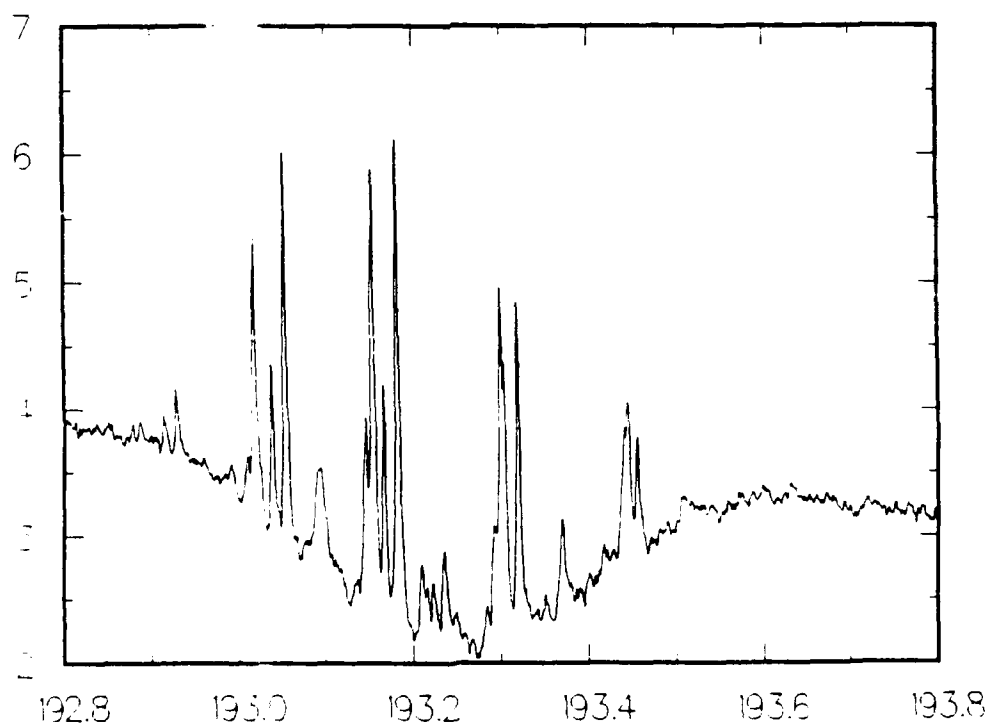


Figure 4. Observed rotationally resolved laser excitation of NO at temperature of 623°K.

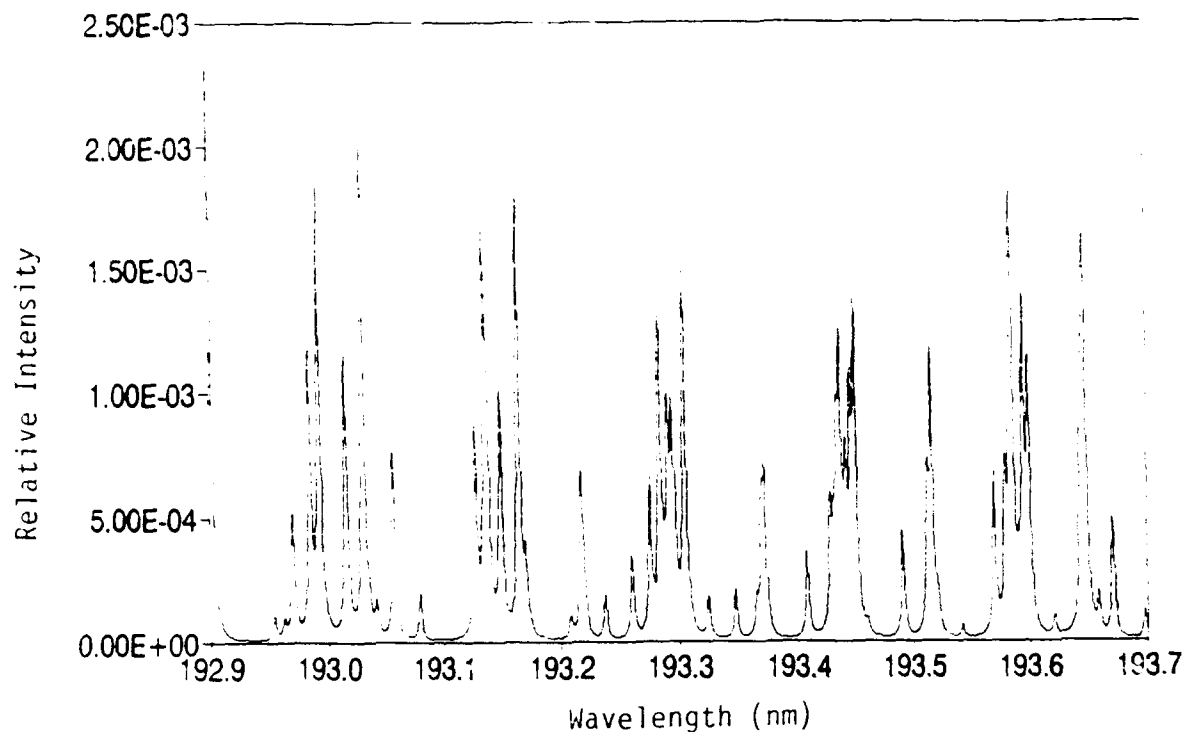


Figure 5. Predicted rotational spectrum of NO at 600°K. The spectrum is the combination of the $B^2\Pi(v'=7) - X^2\Pi(v''=0)$ and $D^2\Sigma(v'=0) - X^2\Pi(v''=1)$ electronic transitions.

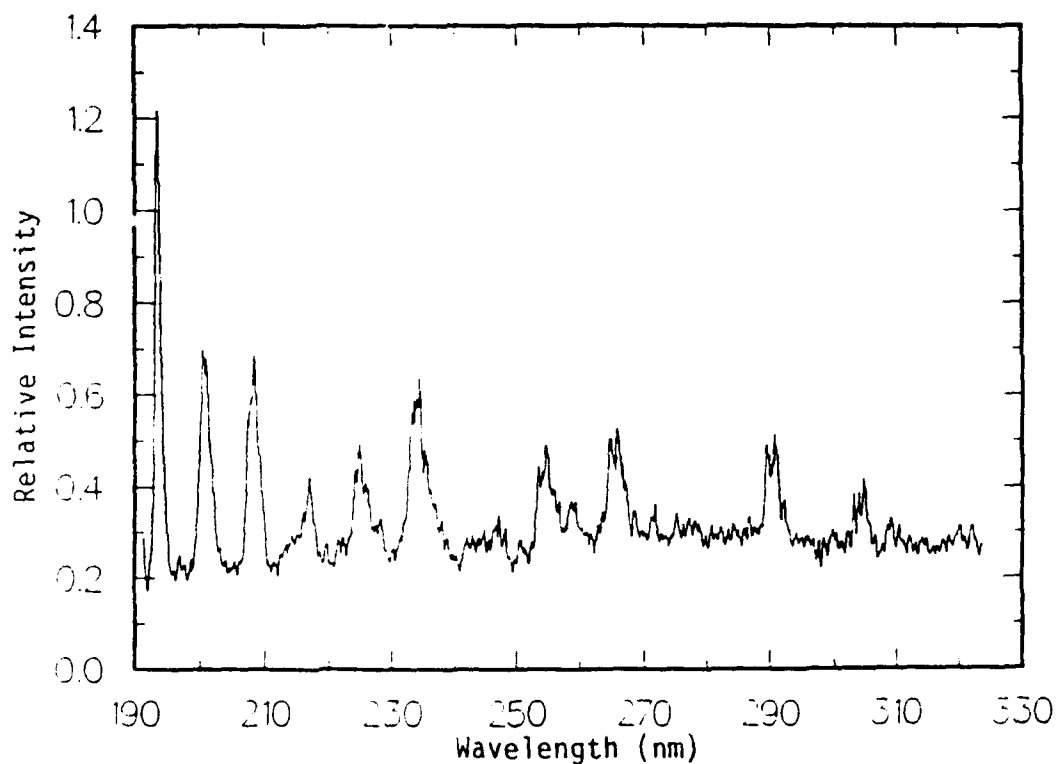


Figure 6. Observed vibrationally resolved fluorescence spectrum of NO at room temperature and a pressure of 1.0 torr with $\lambda = 193.30\text{nm}$.

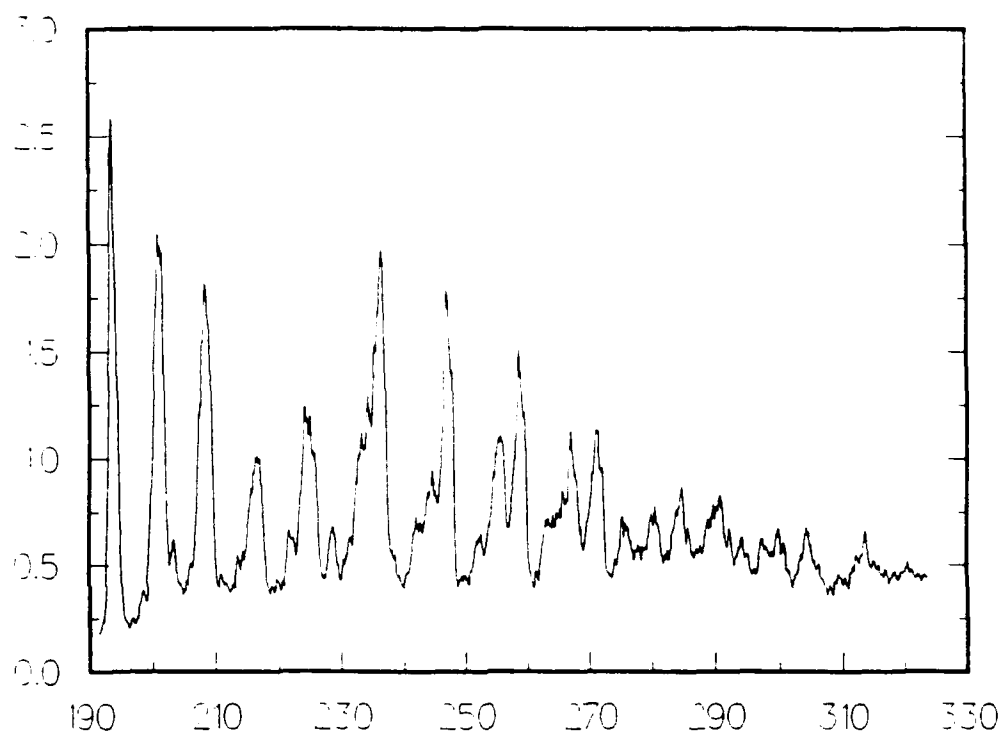


Figure 7. Observed vibrationally resolved fluorescence spectrum of NO at room temperature and a pressure of 10.0 torr with $\lambda = 193.30\text{nm}$.

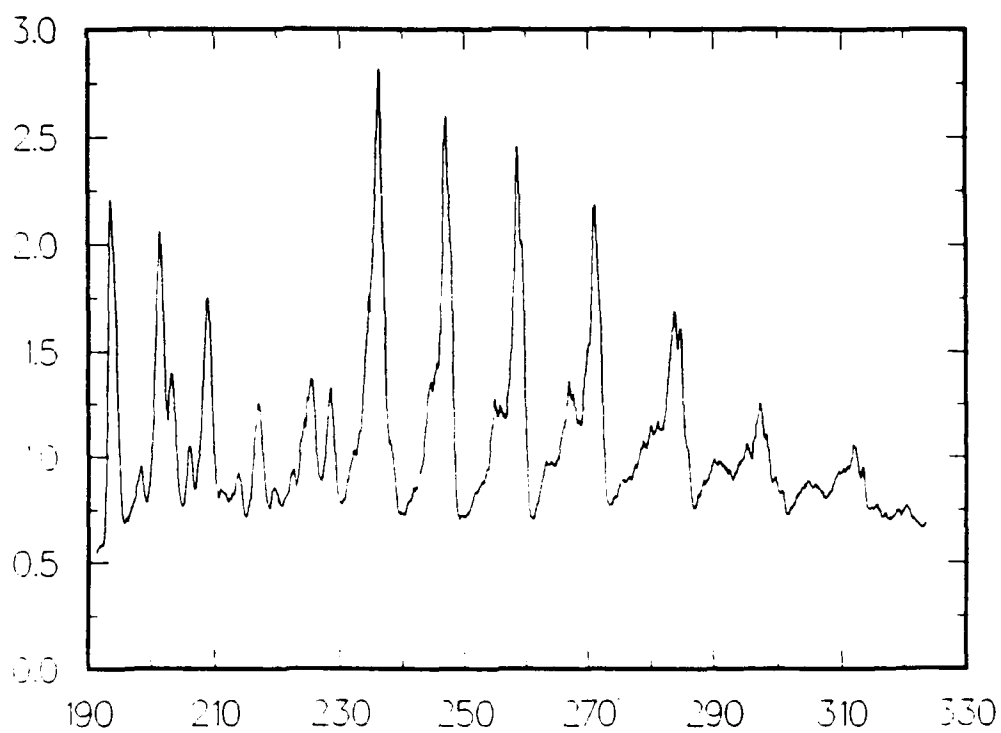


Figure 8. Observed vibrationally resolved fluorescence spectrum of NO at temperature of 623°K and $\lambda = 193.30\text{nm}$.

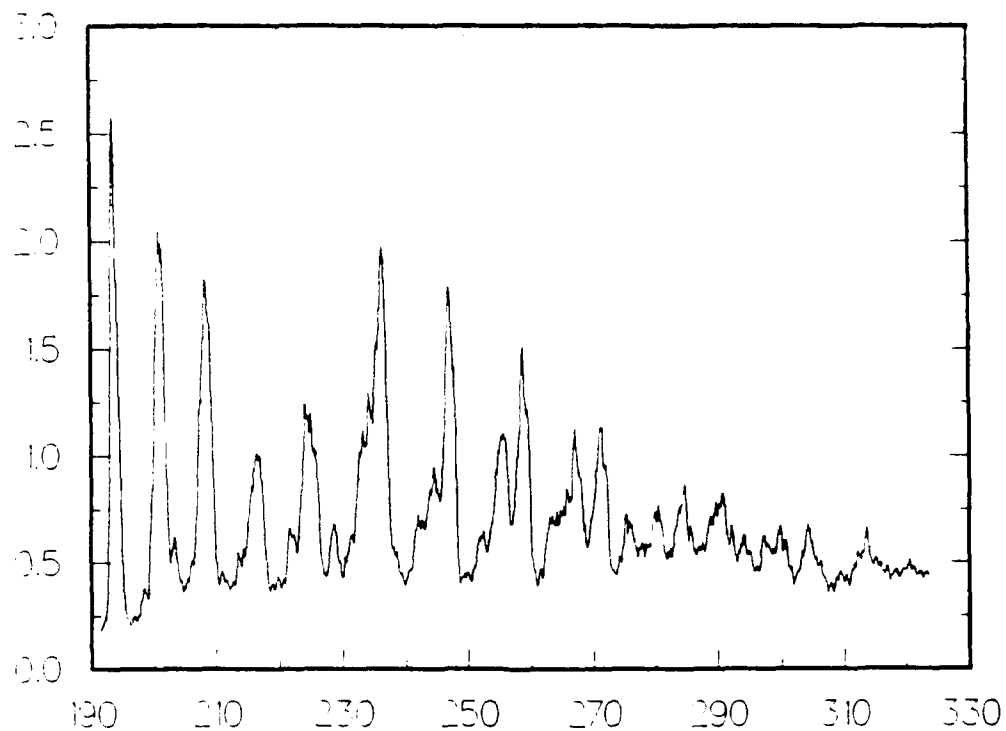


Figure 7. Observed vibrationally resolved fluorescence spectrum of NO at room temperature and a pressure of 10.0 torr with $\lambda = 193.30\text{nm}$.

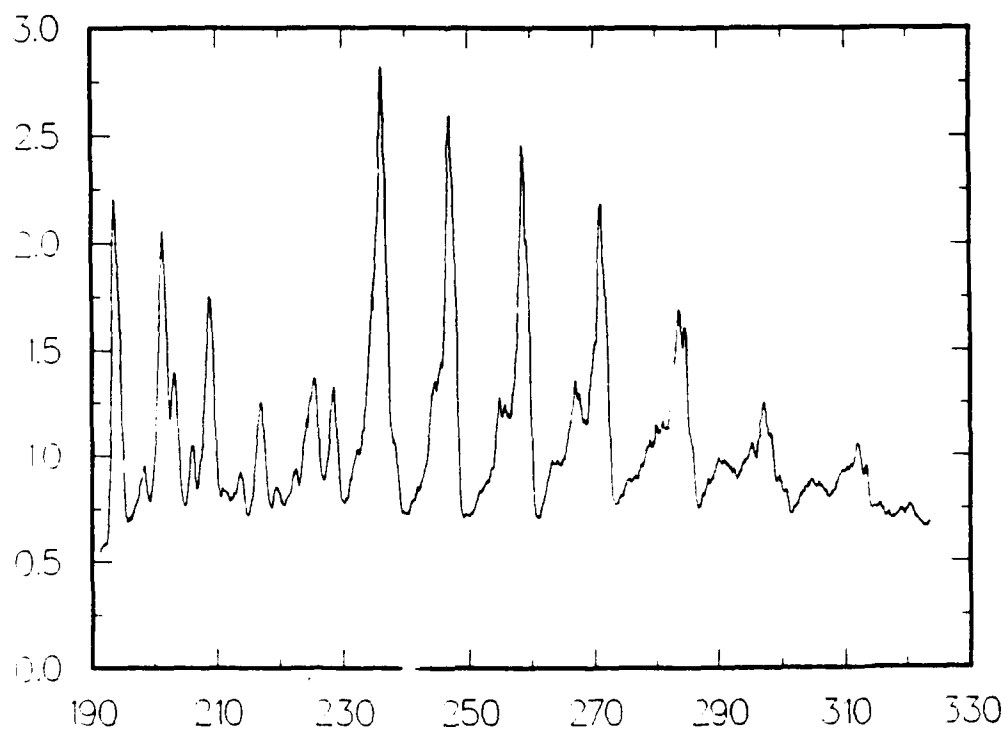


Figure 8. Observed vibrationally resolved fluorescence spectrum of NO at temperature of 623°K and $\lambda = 193.30\text{nm}$.

Table I. Predicted Rotational Spectrum for the $B^2\Pi (\nu' = 7) - X^2\Pi (\nu'' = 0)$ transitions of Nitric Oxide between 192.8 and 193.8 nm.

lower rotational energy level	wavelength (nm)
R11(29 29 1/2)	192.985
P11(26 26 1/2)	192.993
R22(28 27 1/2)	193.015
P22(25 24 1/2)	193.031
R11(30 30 1/2)	193.128
P11(27 27 1/2)	193.136
R22(29 28 1/2)	193.151
P22(26 25 1/2)	193.165
R11(31 31 1/2)	193.273
P11(28 28 1/2)	193.284
R22(30 29 1/2)	193.289
P22(27 26 1/2)	193.305
R11(32 32 1/2)	193.428
R22(31 30 1/2)	193.434
P11(29 29 1/2)	193.437
P22(28 27 1/2)	193.451
R22(32 31 1/2)	193.584
R11(33 33 1/2)	193.586
P11(30 30 1/2)	193.595
P11(30 30 1/2)	193.601

Table II. Assumed Spectroscopic Constants of Nitric Oxide.
Units are in cm^{-1}

$\text{B}^2\Pi(\nu' = 7)$ state	$\text{D}^2\Sigma(\nu' = 0)$ state
Te: 45868.53	Te: 53030.00
We: 1037.45	We: 2323.90
WeXe: 7.47	WeXe: 22.90
WeYe: 0.073	WeYe: 0
Be: 1.031	Be: 1.987
De: 4.9×10^{-6}	De: 5.8×10^{-6}
$\text{X}^2\Pi(\nu'' = 0)$ state	$\text{X}^2\Pi(\nu'' = 1)$ state
Te: 123.16	Te: 122.95
We: 1904.4	We: 1904.4
WeXe: 14.19	WeXe: 14.19
WeYe: 0.024	WeYe: 0.024
Be: 1.696	Be: 1.686
De: 5.3×10^{-6}	De: 5.3×10^{-6}

Table III. Predicted Rotational Spectrum for the $D^2\Sigma(\nu' = 0) - X^2\Pi(\nu'' = 1)$ transitions of Nitric Oxide between 192.8 and 193.8 nm.

lower rotational energy level	wavelength (nm)
Q11(35 35 1/2)+Q21(35 35 1/2)	193.142
Q22(37 36 1/2)+Q12(37 36 1/2)	193.157
R22(31 30 1/2)	193.179
P22(43 42 1/2)+P12(43 42 1/2)	193.209
P11(41 41 1/2)	193.211
S21(23 23 1/2)	193.213
R11(28 28 1/2)+R21(28 28 1/2)	193.222
Q11(34 34 1/2)+Q21(34 34 1/2)	193.223
Q22(36 35 1/2)+Q12(36 35 1/2)	193.246
R22(30 29 1/2)	193.268
P11(40 40 1/2)	193.293
S21(22 22 1/2)	193.293
P22(42 41 1/2)+P12(42 41 1/2)	193.298
R11(27 27 1/2)+R21(27 27 1/2)	193.300
Q11(33 33 1/2)+Q21(33 33 1/2)	193.301
Q22(35 34 1/2)+Q12(35 34 1/2)	193.332
R22(29 28 1/2)	193.354
S21(21 21 1/2)	193.370
P11(39 39 1/2)	193.372
R11(26 26 1/2)+R21(26 26 1/2)	193.375
Q11(32 32 1/2)+Q21(32 32 1/2)	193.378
P22(41 40 1/2)+P12(41 40 1/2)	193.385
Q22(34 33 1/2)+Q12(34 33 1/2)	193.417
R22(28 27 1/2)	193.439
S21(20 20 1/2)	193.445
R11(25 25 1/2)+R21(25 25 1/2)	193.447
P11(38 38 1/2)	193.449
Q11(31 31 1/2)+Q21(31 31 1/2)	193.450
P22(40 39 1/2)+P12(40 39 1/2)	193.469
Q22(33 32 1/2)+Q12(33 32 1/2)	193.498
R11(24 24 1/2)+R21(24 24 1/2)	193.517
S21(19 19 1/2)	193.518
Q11(30 30 1/2)+Q21(30 30 1/2)	193.520
R22(27 26 1/2)	193.521
P11(37 37 1/2)	193.523

SPECTROSCOPIC MONITORING OF
EXHAUST GASES

Final Report
submitted to
Universal Energy Systems, Inc.

R. H. Tipping
Department of Physics & Astronomy
University of Alabama
Tuscaloosa, AL 35487

ABSTRACT

One of the most important molecules that can be used to monitor exhaust gases via spectroscopic means is H_2O . To accomplish this task, one needs to be able to model the spectrum of hot H_2O . Despite the complexity of this problem, we have made substantial progress by deriving a theory for the absorption of radiation by the far wings of spectral lines. Initial validation of these theoretical results have been carried out in the medium infrared spectral region, and the agreement between theory and experiment is generally very good. The relevance of this work to the determination of temperatures is discussed.

I. INTRODUCTION

In order to ascertain information about temperature gradients and/or concentration distributions of molecular species in exhaust gases using spectroscopic techniques, one needs to know (as a minimum) the positions and strengths of the radiative transitions involved. For many species this information is not available directly from laboratory measurements because of the conditions under which most experiments are performed, viz. relatively low temperatures. The extrapolation of this laboratory data to the high temperatures existing in exhaust gases is not a simple problem, especially for polyatomic molecules, because of the existence of both highly excited rotational levels and a large number of hot bands that are present in the spectra of hot gases. In the present report we shall describe briefly some progress that has been made in compiling a spectroscopic data base that would enable one to model the spectra of several exhaust species.

One of the main molecules of interest is H_2O . However, because of the complexity of the problem of modeling the spectrum of hot water vapor, an adequate data base is not currently possible, although substantial progress has been made as will be discussed later. On the other hand, accurate frequency and intensity data for several isotopic species of CO were calculated; these would enable one to model the vibration-rotational spectrum of CO up to several thousand degrees. This data was already made available to my research colleagues at Arnold Development Center, and it will be included in the HITEMP

data base currently being assembled at the Air Force Geophysics Laboratory, Hanscom Air Force Base. A similar data base for the Schuman-Runge electronic transitions of O_2 has been completed and made available to the workers at Arnold Development Center.

In view of the above, rather than reviewing this work on diatomic molecules that has been completed, we will focus mainly on the progress that has been made on the problem of hot water vapor. In the following, we will first describe work done in collaboration with a graduate student (Q. Ma) on the calculation of atmospherically important "water continuum". The extrapolation to high temperatures and its relevance to the monitoring of exhaust gases will also be briefly discussed. Then we will discuss some recently completed experimental work and theoretical analysis carried out in collaboration with J.-M. Hartmann of the Laboratoire d'Energetique Moleculaire et Macroscopique Combustion, l'Ecole Centrale in Paris, France. Together this body of work constitutes an important step towards the goal of a quantitative understanding of the spectrum of hot water vapor.

II. THE WATER SPECTRUM

A. The "Water Continuum"

The continuum absorption of radiation in the infrared windows of the Earth's atmosphere has been known for a long time.¹ The history of this topic, together with many references to earlier work, has been reviewed recently by a number of authors.²⁻⁵ There is nearly unanimous agreement on the

temperature dependence (strong, negative), but there is considerable disagreement as to its magnitude and the physical mechanism responsible for the absorption.⁶ Water dimers, collision-induced absorption, and the superposition of the far wings of collisionally broadened lines have all been proposed as possible candidates. While the first two mechanisms are undoubtedly present and contribute to the absorption (in varying amounts in the different spectral regions), it is now generally accepted that the major contribution to the absorption coefficient in the spectral region between 300-1100 cm^{-1} is due to the wings of the strong self-broadened pure rotational transitions of water.

We have derived a theory valid in the far wings of water transitions that has as input only well-known molecular parameters; this theory enables us to calculate the absorption due to overlapping of far wings from the millimeter region through the infrared.^{7,8} The results of such a calculation are shown in Figure 1 for three temperatures: $T = 296$ K (solid curve), $T = 338$ K (dashed curve), and $T = 430$ K (dotted curve). We have also carried out similar calculations for higher temperatures ($T = 760$ K) more relevant to exhaust gases. Obviously, the temperature-dependence of these results could be used in conjunction with broadband measurements in the regions between the vibration-rotational bands (eq. near 1000 cm^{-1} , 2800 cm^{-1} , etc.) to determine the temperature of hot water vapor. More work is needed, however, to verify our results using laboratory data, before such a quantitative comparison could be

made. Initial attempts in the medium, infrared region will now be described.

B. Medium Infrared Measurements and Analysis

Experimental measurements have been made in the high frequency wings of the ν_2 -band and the $(\nu_1, 2\nu_2, \nu_3)$ -triad for temperatures and pressures in the 500-900 K and 0-70 atmospheres ranges. Because these results will be published in the near future,⁹ we will not give the experimental details here, rather we will concentrate on the results. In Figure 2 we present experimental results for the continuum absorption (circles) in the 400-900 cm^{-1} frequency range. The dotted curve gives the results calculated from our theory to be compared with the dashed curve which one would obtain using the conventional Lorentz lineshape; clearly the new results are a vast improvement and lead to excellent agreement at high frequencies where the assumptions in the theory^{7,8} are valid. Also shown is an improved fit (solid line) based partly on the laboratory data. Similar results have been obtained in other spectral regions.⁹

III. DISCUSSION AND CONCLUSIONS

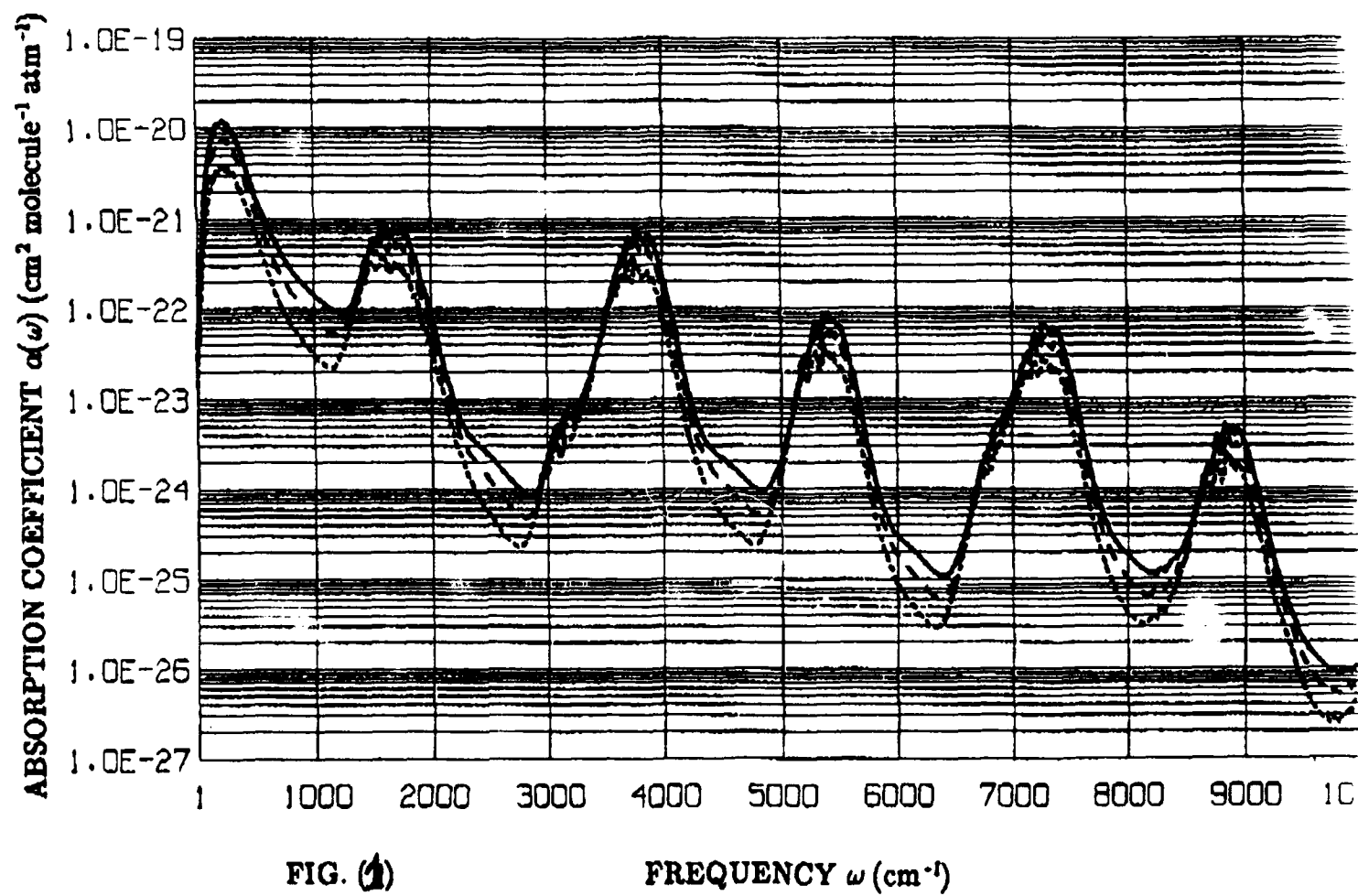
Substantial progress has been made in deriving a theory for the calculation of the absorption by hot water vapor, primarily in the regions between the strong vibration-rotational bands. We have validated the theoretical calculations by comparison with new laboratory data obtained at elevated temperatures and pressures. By such comparisons, one can improve the results

and model quite accurately the absorption spectrum. Because of the strong temperature dependence of the absorption, one could in principle obtain temperatures from the measured absorption.

Despite the success of our results,⁷⁻⁹ much work still remains; in particular, the calculation of the absorption due to individual vibration-rotational transitions that would enable one to obtain density and temperature data from high resolution measurements as opposed to broad-band ones. Only when such results are available can they be combined with the absorption arising from the far wings in order to predict quantitatively the spectrum of hot water over a wide range of densities and temperatures.

REFERENCES

1. W. M. Elsasser, *Astrophys. J.* 87, 497 (1938).
2. S. A. Clough, F. X. Kneizys and R. W. Davies, *Atmos. Res.* 23, 229 (1989).
3. W. B. Grant, *Appl. Opt.* 29, 451 (1990).
4. M. E. Thomas, *Infrared Phys.* 30, 161 (1990).
5. P. Varanasi, *SPIE Proc.* 928, 213 (1988).
6. Atmospheric Water Vapor, ed. by A Deepak, T. D. Wilkerson and L. H. Runke, Academic Press, New York, 1980.
7. Q. Ma and R. H. Tipping, *J. Chem. Phys.* 93, 6127 (1990).
8. Q. Ma and R. H. Tipping, *J. Chem. Phys.* 93, 7066 (1990).
9. J. -M. Hartmann, M. Y. Perrin, Q. Ma and R. H. Tipping, *J. Chem. Phys.*, submitted for publication.



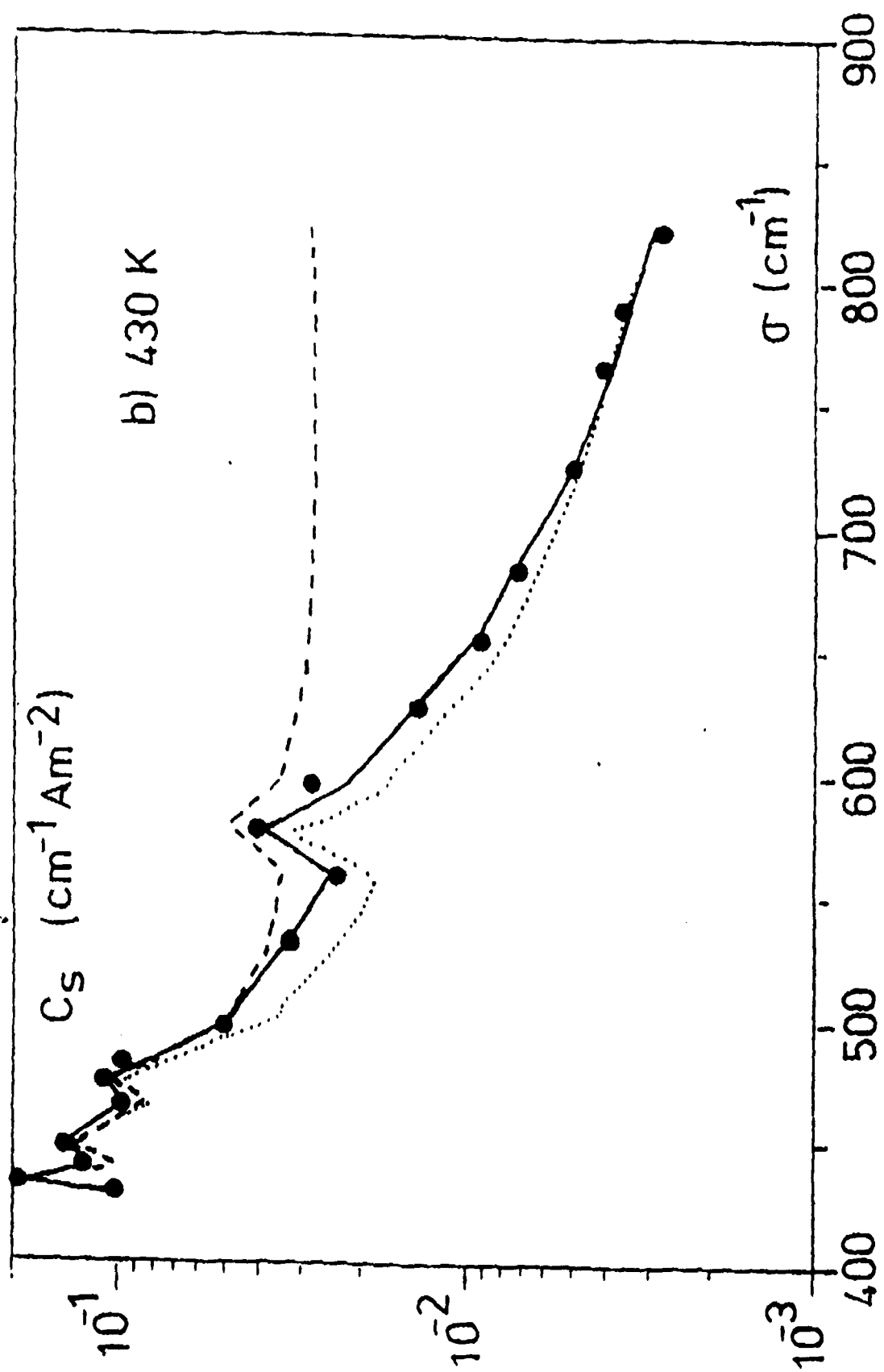


FIGURE 2.

Transient Analysis of Parallel Distributed Structurally Adaptive Signal Processing Systems

Final Report Submitted to:

USAF-UES Summer Faculty Research Program

Contract Number:

F49620-88-C-0053/SB5881-0378

Sponsored by the

Air Force Office of Scientific Research

Conducted by:

Universal Energy Systems, Inc.

Submitted by

D. Mitchell Wilkes
Vanderbilt University
Department of Electrical Engineering
Box 1649 Station B
Nashville, TN 37235
(615) 343 - 6016

December 1990

Abstract

This report describes the research accomplished under the USAF-UES Summer Research Program (SRP) mini-grant awarded for the period 1 January 1990 to 31 December 1990. The main goal of this research was the analysis of the dynamic transient properties of a simple class of structurally adaptive signal processing systems. The work performed on this grant shows that even a very simple type of system adaptation (one so simple that it would not ordinarily be considered a structural adaptation) will usually require a dynamic modification of the processing structure in order to avoid an undesirable transient response to the change. Additionally, it is possible to predict when an undesirable transient will occur, and thus to perform the structural modification to avoid it.

1 Acknowledgements

I would like to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsoring this research. I would also like to thank Universal Energy Systems for their help in the administrative aspects of the program. Many thanks also to Arnold Engineering Development Center(AEDC)/DOTR for their help and gracious use of their equipment.

This work is also heavily dependent on the work of Capt. Ted Bapty for his help and guidance at AEDC and Andrew Koffman for his help in performing the transient analysis of structural adaptation. Also, I must thank Dr. Janos Sztipanovits for his guidance in leading me into the area of structurally adaptive signal processing and for producing his brainchild, MULTIGRAPH.

2 Introduction

A complete, parallel, fully-integrated programming and execution environment, the Multigraph Programming and Execution Environment (MPEE), for structurally adaptive signal processing systems has been developed at Vanderbilt University [1]–[15]. This environment provides a user-friendly environment for designing, programming, and executing a signal processing system having the ability to modify its own processing structure while it is running. This represents a significant generalization of the present level of adaptive systems which merely adjust parameters within a fixed structure.

It is possible to design and implement structurally adaptive systems with existing programming languages such as C or Fortran, but the Multigraph system has several important advantages:

- It is a generic environment for designing and implementing processing systems that are representable in a block diagram or signal flow graph form. Therefore, the user is able to organize his thoughts in a natural way around the flow of the signal through the system, rather than in the unnatural methods of subroutine and function calls so prevalent in standard programming languages.
- Also, in an ordinary programming environment, the user must provide his own mechanism for deciding which processing tasks must be performed and in what order they are to be performed (i.e., the scheduling of processing tasks in the block diagram or signal flow graph). In the MPEE scheduling is simplified, since the final execution environment has a macro-dataflow computational structure, and thus closely resembles the signal flow graph of the processing system. The result is that the execution environment can dynamically determine the proper scheduling of the processing tasks on its own, and free the user from this tedious task. Actually, in a structurally adaptive system this task is more than just tedious, it is critical. Since the overall application system can change its processing graph structure, the schedule will change during run-time. This means that the dynamic scheduling capability of the execution environment is crucial for the proper operation of the overall system. This capability would generally be difficult to implement with standard programming languages, and would require the generation of a dynamic scheduling system for each new application program.
- In addition, the macro-dataflow computational model realized by the execution environment lends itself very well to parallel multiprocessor implementations, with the result that such implementations have been successfully tested.
- An additional advantage of the MPEE is its hierarchical approach to designing large-scale systems. The design and implementation of flexible large-scale signal processing

systems can become so unwieldy that standard programming languages quickly become impractical. The MPEE is well-suited to large-scale systems, because higher-level graphic-oriented modeling facilities for signal processing structures are provided in the programming environment. Essentially, large-scale signal processing systems can be designed by "drawing pictures" on the computer monitor.

The MPEE provides a nearly ideal environment for designing and implementing structurally adaptive signal processing systems, but more work needs to be done to analyze the signal and system theoretical behavior of this class of adaptive system. In particular, the transient behavior of such systems must be analyzed. The work reported herein is a step toward addressing this need.

3 Research Goals

The main goal of this effort was to investigate the signal and system theoretical properties of structurally adaptive signal processing systems with special attention to the characterization of their dynamic, transient behavior. Since the designation, structurally adaptive system (SAS), is much too general for any in-depth analysis, the method to be used to study this type of system is to define a pool of problems representative of different kinds of SAS's and start analyzing these examples.

In order to make use of the powerful mathematical tools that exist for analyzing linear time-invariant systems, the scope of this effort was limited to studying the dynamics of reconfiguration for linear filter structures. The pool of experimental systems for this study that fell into the category of linear time-invariant (LTI) filters included the following structures:

- Recursive transformations: This structure looks quite promising for implementing a broad class of SAS's. Advantages include finite transient time, a simple initialization process, and a simple method of modification [17, 18].
- Direct form I and II: These structures are easily initialized by keeping track of past values of input and output. On the other hand, they are not well behaved under coefficient quantization, and stability is not assured in a simple manner (the Schur-Cohn stability criterion must be used).
- Lattice filters: These filters are formed of a cascade of sections having identical structure. As such, the order of the filter is easily increased or decreased. Also, they are well behaved under coefficient quantization and stability is easily assured by checking the magnitude of the reflection coefficients.

- Wave digital filters: These have the property that they are easily designed according to frequency domain specifications, and they have robust numerical properties [16].
- Cascade filter structures: Cascades of second-order sections have several desirable qualities. They are relatively insensitive to coefficient quantization and stability is easily assured. Also, this type of structure directly represents pole/zero information.
- Parallel filter structures: As with cascade structures, parallel combinations of second-order sections are relatively insensitive to coefficient quantization and stability is easily assured. In contrast, however, while pole information is directly represented by these sections, the locations of the zeroes are not as easily identified.
- Composite structures: These are larger signal processing systems composed of interconnections of linear subsystems.

Clearly, the study of all or even most of these structures is well beyond the scope of a twelve month effort, so the direct form and lattice structures were chosen for investigation and comparison. Through these, especially the direct form structures, it was possible to obtain results that are generally applicable to LTI systems that are subject to dynamic reconfiguration while running.

Further reduction in scope was necessary. The set of all possible structural changes that may be implemented on a running filter is enormous. Therefore, the simplest possible change, i.e., merely changing the parameters of a running filter, was chosen for the initial study. Such a small change requires only that the parameters (or coefficients) of the filter be changed without actually changing the processing structure of the filter. Strictly speaking this is not a true structural adaptation, since the processing structure does not change. However, this simple modification proved to require dynamic structural modification in order to deal with a potentially very undesirable transient response. Thus, even the simplest modification proved the need for the ability to implement dynamic structural modifications. The analysis performed in this effort focused on characterizing the transient response, predicting when it would be unacceptably large, and exploring methods for dealing with the transient.

4 Research Performed

The work reported in the section is described in greater detail in [20] and [21].

4.1 Transient Response of a Pole-Zero System

To determine the conditions under which undesirably large transient behavior can occur at the output of a system, and what steps if any can be taken to correct these conditions, it

is helpful to perform a mathematical analysis of the system. Practical digital filters can be represented in the Z-domain as a ratio of two polynomials in z , thus having a finite number of poles and zeros. In this effort, a digital filter experiencing an abrupt change in its coefficients was considered to be a pole-zero system with non-zero initial conditions. Thus a dynamically reconfigurable digital filter will have a non-zero state after the initial point of reconfiguration, i.e., the new filter will have an initial state that corresponds to the final state (and thus relates to the frequency characteristics) of the old filter. Therefore, if the coefficients are abruptly changed, compatibility with the state of the old filter is an important issue.

The following equations describe the Z-domain response of the filter after the point of reconfiguration. It is assumed that an abrupt change of the coefficients occurred just prior to time $n = 0$. Let $Y^+(z)$ be the one-sided Z-transform of the nonnegative time output signal $y(n)$, $X^+(z)$ be the one-sided Z-transform of the nonnegative time input signal $x(n)$, the a_k 's be the denominator coefficients of the new filter, and the b_k 's be the new numerator coefficients. If the new filter obeys the following difference equation (for $n \geq 0$)

$$y(n) = - \sum_{k=1}^N a_k y(n-k) + \sum_{k=0}^M b_k x(n-k) \quad (1)$$

then the one-sided Z-transform yields [22],

$$Y^+(z) = - \sum_{k=1}^N a_k z^{-k} (Y^+(z)) + \sum_{n=1}^k y(-n) z^n + \sum_{k=0}^M b_k z^{-k} (X^+(z)) + \sum_{n=1}^k x(-n) z^n. \quad (2)$$

Simplifying, it is straightforward to show that

$$\begin{aligned} Y^+(z) &= \frac{B(z)}{A(z)} X^+(z) + \frac{\sum_{k=0}^M b_k z^{-k} (\sum_{n=1}^k x(-n) z^n)}{A(z)} - \frac{\sum_{k=1}^N a_k z^{-k} (\sum_{n=1}^k y(-n) z^n)}{A(z)} \\ &= \frac{B(z)}{A(z)} X^+(z) + \frac{N_{0x}(z)}{A(z)} + \frac{N_{0y}(z)}{A(z)} \\ &= H(z) X^+(z) + Y_{0x}(z) + Y_{0y}(z). \end{aligned} \quad (3)$$

where $A(z) = 1 + \sum_{k=1}^N a_k z^{-k}$ and $B(z) = \sum_{k=0}^M b_k z^{-k}$. $H(z)$ is the new transfer function of the reconfigured filter. $Y_{0x}(z)$ and $Y_{0y}(z)$ are the transient responses due to the states from the past inputs and past outputs of the system, respectively.

It can be shown that the term $N_{0y}(z)$ depends on the initial states of the output and the denominator of $H(z)$. The numerator of $Y_{0y}(z)$, the term derived from the initial states of the output at time $n = 0$, can be decomposed as follows:

$$\begin{aligned} N_{0y}(z) &= - \sum_{k=1}^N a_k z^{-k} \left(\sum_{n=1}^k y(-n) z^n \right) \\ &= -[w_2(0) + w_2(1)z^{-1} + \dots + w_2(N-2)z^{-(N-2)} + w_2(N-1)z^{-(N-1)}] \quad (4) \\ &= -W_2(z) \end{aligned}$$

where the $w_2(n)$ terms may be computed in the following way. Let $y_0(n) = \sum_{i=1}^N y(-i)\delta(n+i)$ (where $\delta(n)$ is the unit impulse), $a(n) = \sum_{i=1}^N a_i\delta(n-i)$, and $\tilde{w}_2(n) = a(n) * y_0(n)$, then, $w_2(n)$ may be found from

$$w_2(n) = \tilde{w}_2(n)u(n) \quad (5)$$

where $u(n)$ is the unit step function.

4.2 Analysis of the Transient

The analysis of the transient response of the system will be based on the preceding decompositions of the system components. From Equation (3), we see that the response of a filter, after the point of abrupt coefficient change, is due to three components. The first term, $H(z)X^+(z)$, is simply the response of the new filter to the new input data. Provided the new filter is stable, this term should not produce an undesirably large transient. The other two components, $Y_{0x}(z)$ and $Y_{0y}(z)$, are the responses resulting from the past states of the input and output of the old filter, respectively. These final states could be set to zero after a change of coefficients, but this would cause a discontinuity in the system response, which may or may not be acceptable. The past state of the input is not likely to cause a large transient [20]. The coefficient change does, however, cause the recursive section of the new filter to be driven by the past output states of the old filter. The past states of the output are the new filter's link to the old filter characteristics, and incompatibilities between the old and new filters will be manifested in the term $Y_{0y}(z)$.

Another perspective can be gained from looking at the numerator of $Y_{0y}(z)$ in terms of the filtering operations mentioned above. Let $Y_0(z)$, $A(z)$, $\tilde{W}_2(z)$, $W_2(z)$, and $U(z)$ be the Z-transforms of $y_0(n)$, $a(n)$, $\tilde{w}_2(n)$, $w_2(n)$, and $u(n)$, respectively. Then from the decomposition of $Y_{0y}(z)$, it can be seen that

$$Y_{0y}(z) = \frac{N_{0y}(z)}{A(z)} = -\frac{W_2(z)}{A(z)}, \quad (6)$$

which by partial fraction expansion can be transformed into (assuming simple poles)

$$Y_{0y}(z) = \sum_{i=1}^N \frac{A_i z}{z - p_i} = \sum_{i=1}^N \frac{A_i}{1 - p_i z^{-1}}, \quad (7)$$

The partial fraction expansion coefficients can be written as

$$A_i = - \left. \frac{W_2(z)(1 - p_i z^{-1})}{A(z)} \right|_{z=p_i}. \quad (8)$$

In order to consider the effects these coefficients have on the transient response, Equation (8) may be considered as the product of two terms:

$$A_i = \left[\frac{1 - p_i z^{-1}}{A(z)} \right]_{z=p_i} [W_2(z)]_{z=p_i}. \quad (9)$$

The numerator of the first term cancels the i^{th} root of $A(z)$, leaving an all-pole expression. Since the expression is evaluated at $z = p_i$, the first term will grow large if any of the other poles lie close to the i^{th} pole, or if the filter is of high order and several poles lie relatively near the i^{th} pole. Either of these two situations can lead to an undesirably large transient behavior. The magnitude of A_i will also be affected by the magnitude of the second term in Equation (9). If the second term is small enough, it may cancel a large value in the first term. Otherwise, a large transient may result. Now we consider the magnitude of $W_2(p_i)$.

Since, $w_2(n)$ equals the product of $\tilde{w}_2(n) (= a(n) * y_0(n))$ and $u(n)$, the second term in the expression for A_i can be evaluated via complex convolution of Z-transforms. Taking $p_i = re^{j\phi}$ and evaluating the convolution integral on a circle having a radius equal to that of p_i results in

$$W_2(p_i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} U(e^{j(\phi-\theta)}) A(re^{j\theta}) Y_0(re^{j\theta}) d\theta. \quad (10)$$

However, a simplification results from examining $U(e^{j(\phi-\theta)}) = \frac{1}{1-e^{-j(\phi-\theta)}}$ and

$$A(re^{j\theta}) = (1 - p_1 r^{-1} e^{-j\theta}) \dots (1 - e^{j(\phi-\theta)}) \dots (1 - p_N r^{-1} e^{-j\theta}). \quad (11)$$

Multiplying $U(e^{j(\phi-\theta)})$ and $A(re^{j\theta})$, we see that the i^{th} term of $A(re^{j\theta})$ combines with $U(e^{j(\phi-\theta)})$ in the following manner:

$$\frac{1 - e^{j(\phi-\theta)}}{1 - e^{-j(\phi-\theta)}} = -e^{j(\phi-\theta)}. \quad (12)$$

Thus, the product simplifies to

$$U(e^{j(\phi-\theta)}) A(re^{j\theta}) = -e^{j(\phi-\theta)} \prod_{k \neq i} (1 - p_k r^{-1} e^{-j\theta}) = -e^{j(\phi-\theta)} Q(re^{j\theta}), \quad (13)$$

which results in

$$W_2(p_i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} -e^{j(\phi-\theta)} Q(re^{j\theta}) Y_0(re^{j\theta}) d\theta. \quad (14)$$

For each pole of the filter, $W_2(p_i)$ is a factor of the i^{th} partial fraction expansion coefficient. Since the transient magnitude is directly proportional to the coefficient, it is also directly proportional to $W_2(p_i)$. The magnitude of this integral is dependent on the relative energies of the two terms. It is desirable to have the energy or "high regions" of $Y_0(re^{j\theta})$ fall in the vicinity of zeros of $Q(re^{j\theta})$, which are at the locations of the poles of the filter (roots of $A(re^{j\theta})$). Such an overlap will reduce the magnitude of the transient. Energy in $Y_0(re^{j\theta})$ that is amplified by $Q(re^{j\theta})$ will tend to increase the size of the transient.

These points are demonstrated in Example 1 which examines the components found in the convolution integral for two situations.

Example 1: A filter is redesigned at time zero to a fourth-order narrowband bandpass filter (with closely spaced poles at $0.1 \pm j0.96$ and $0.0999 \pm j0.9599$) driven only by past outputs of an unknown filter. The main interest here is the location of the energy contained in the spectrum of $|Y_0(re^{j\theta})|$. We will consider the cases in which the dominant energy of $|Y_0(re^{j\theta})|$ is contained essentially near the i_{th} root of $|A(re^{j\theta})|$ (Case 1) and away from the i_{th} root of $|A(re^{j\theta})|$ (Case 2). We can see in Figure 1 the plots of $|Q(re^{j\theta})|$ and $|Y_0(re^{j\theta})|$ for Case 1. Note that the relative maxima of $|Q(re^{j\theta})|$ match up with the relative minima of $|Y_0(re^{j\theta})|$, and vice versa, disallowing any significant amplification of energy. The resulting transient response of the filter is shown in Figure 2. Figure 3 shows the plots of $|Q(re^{j\theta})|$ and $|Y_0(re^{j\theta})|$ for Case 2. In this case, the maxima of the two terms are closely aligned. The transient response of the filter for Case 2 is shown in Figure 4. This transient is much larger than that of Case 1.

4.3 Some Possible Reconfiguration Methods

Since reconfiguring (or changing the coefficients of) a digital filter by abruptly changing the filter coefficients can produce a large transient response, certain situations call for other more gradual approaches to the reconfiguration problem. This section considers other methods for changing the filter coefficients. The methods examined come in two basic forms: interpolation methods (which incrementally change the coefficients from the old filter into those of the new filter) and dynamic parallel implementation methods (which switch filters by running the new filter in parallel with the old filter for a time).

4.3.1 Interpolation Methods

Possibly the simplest choice for interpolating between filters is a linear interpolation of the direct form filter coefficients. For a filter with a transfer function of the form $H(z) = B(z)/A(z)$ we define a time period, T , over which the coefficient change will occur. If the change begins at time $n = n_0$, then the end of the change period will be $n = n_0 + T$. Thus, the filter coefficients will be interpolated linearly according to

$$A_n(z) = \frac{n - n_0}{T} A_{new}(z) + \frac{n_0 + T - n}{T} A_{old}(z), \quad n_0 \leq n \leq n_0 + T. \quad (15)$$

The coefficients of $B(z)$ are interpolated in the same manner. One potential problem is that as the coefficients of $A(z)$ are interpolated, the corresponding poles cannot be guaranteed to remain within the unit circle, thus stability cannot be guaranteed [20].

The problem of instability can be eliminated by interpolating between the pole locations of the old and new stable filters in a manner similar to the previous interpolation method. The numerator can be changed by interpolating filter zeros in the same fashion or by using the coefficient interpolation described above. While this method prevents the system from

becoming unstable (provided the initial and final filters are stable), there may be another factor contributing to unacceptable transient behavior. Pole-zero systems exhibit a sensitivity to shifts in pole locations which can cause the frequency response of the filter being interpolated to develop a high gain or otherwise become unacceptable in such a way as to produce an undesirable transient [20].

Another method for trying to preserve stability is the interpolation of lattice parameters. Instability should not be a factor, since a lattice structure with reflection coefficients having magnitude less than unity will always be stable. Therefore, given initial and final filters that are stable, all intermediate filters will be stable. However, unacceptable transient behavior can still exist because of a sensitivity to shifts in pole locations [20]. In summary, these three interpolation methods were examined via experiments in [20] and found to give unacceptably large transients.

4.3.2 A Dynamic Parallel Implementation Method

Clearly, implementing a change of the coefficients of a filter by directly modifying the coefficients, either abruptly or gradually, can produce undesirable transient behavior. It may be possible to determine a sequence of filter coefficients that will not produce an undesirable transient, but finding that sequence is not a trivial task. A simpler and more predictable technique for achieving a well-behaved change of the filter coefficients proceeds as follows. Build a new filter in parallel with the old one, forcing the new filter's initial states to be all zeros, then perform a time-varying weighted sum of the outputs of the two filters. The time-varying weighted sum consists of linearly changing the values of the amplifiers at the inputs to a summer over the switching time, T . The weight associated with the old filter will change linearly from one to zero, and that of the new filter will change linearly from zero to one. The output of the filter, $y(n)$, during the switching process is

$$y(n) = k_1(n)[h_{old}(n) * x(n)] + k_2(n)[h_{new}(n) * x(n)] \quad (16)$$

where $k_1 = \frac{n_0+T-n}{T}$ and $k_2 = \frac{n-n_0}{T}$ for $n_0 \leq n \leq n_0 + T$. Assuming both filters are properly behaved, no undesirable transient will be produced. The implementation form for this method is shown in Figure 5. Figure 5(a) represents the original filter prior to change. Figure 5(b) shows the multipliers that are varied during the tapered switch. Figure 5(c) is simply the new filter after switching. A very similar dynamic parallel implementation method, the overlapping sum method, is given in [20]. Example 2 shows a tapered switch applied to a bandpass filter.

Example 2: A linearly tapered switch implementation of a bandpass filter redesign has been performed over the time interval $n = 301$ to $n = 400$. The transient response produced is well-behaved as shown in Figure 6.

4.4 Determination of Transient Size

The purpose of determining the size of the transient prior to implementation of the filter change is to determine which type of reconfiguration to implement. If transient behavior never becomes undesirably large, an abrupt coefficient change may be perfectly acceptable. However, if it can be determined that the transient will grow large, then some other type of change technique will be necessary.

4.4.1 Exact Calculation of Transient

Section 4.2 showed that given stable old and new filters and bounded input, any undesirably large transient behavior occurring at the time of filter change will be largely caused by the incompatibility between the initial output state (from the old filter) and the new filter coefficients. The transient due to the initial output states can be determined by creating a filter with transfer function $\frac{N_{0y}(z)}{A(z)}$, driving this filter with an impulse, and searching for the maximum. This computational method obtains the output state transient, and requires only knowledge of the filter coefficients and initial output state.

4.4.2 Estimate of Transient

An estimate of the maximum transient of $y_{0y}(n)$ can be obtained by sampling its Z-transform, $Y_{0y}(z)$, N times around the unit circle and then applying an inverse fast Fourier transform (IFFT).

$$\hat{Y}_{0y}(k) = Y_{0y}(z)|_{z=\exp j\frac{2\pi}{N}k} \quad k = 0, 1, \dots, N-1$$

The transient estimate is then

$$\hat{y}_{0y}(n) = IFFT\{\hat{Y}_{0y}(k)\}. \quad (17)$$

The largest transient value would then be estimated by $\max_n \{|\hat{y}_{0y}(n)|\}$. This estimate can be both fast and have predictable (fixed) computational cost. However, sampling the Z-transform, $Y_{0y}(z)$, produces an aliased inverse transform that may require N to be large to diminish the aliasing. N may not have to be large to obtain a useable estimate, and practical results have been obtained for values as small as $N = 8$ [20].

4.4.3 Transient Bounds

Direct Time-Domain Bound: Given that the filter implemented after reconfiguration is a stable filter, all the poles will lie inside the unit circle and, therefore, have magnitude less than one. Thus, we see from the time-domain transient response that the largest value of the

transient will always be less than the sum of the magnitudes of the partial fraction expansion coefficients, A_i . Therefore, one bound for the transient response is given by

$$\max_n |y_{0y}(n)| \leq \sum_{i=1}^N |A_i|. \quad (18)$$

To obtain this bound, the roots of the denominator polynomial of the filter will need to be determined. The computational cost of this operation (if the roots are unknown) and the closeness of the bound to the actual transient maximum are both unknown.

Fourier Transform Bound: Another bound on the transient can be obtained by sampling the Fourier transform of $y_{0y}(n)$ from its Z-transform, $Y_{0y}(z)$, N times around the unit circle and taking the inverse Fourier transform. The discrete Fourier transform, $\hat{Y}_{0y}(k)$, which is an estimate of the Fourier transform, $Y_{0y}(e^{j\omega})$, will be given by

$$\hat{Y}_{0y}(k) = Y_{0y}(z)|_{z=\exp j\frac{2\pi}{N}k}, \quad k = 0, 1, \dots, N-1. \quad (19)$$

Applying the inverse Fourier transform, taking the absolute value of both sides, and applying the triangle inequality results in

$$|\hat{y}_{0y}(n)| \leq \frac{1}{N} \sum_{k=0}^{N-1} |\hat{Y}_{0y}(k)| \quad (20)$$

for all n . Comparing the cost of obtaining this bound with the costs of obtaining the time-domain bound, the estimate, and the exact calculation of the transient, this bound may be the least expensive to compute. No rooting of the filter denominator polynomial is required, no search for a maximum is required, and no FFT is needed. The accuracy of this bound will, however, depend on the number of samples N , and in practice it has been observed that this bound is not close to the actual maximum unless N is chosen to be large (e.g., $N = 1024$ or 2048).

Comparison of Bound and Estimate Accuracy: Bound and estimate data for the transient term due to the initial output state have been taken for a fourth-order narrowband bandpass filter (with poles at $0.1 \pm j0.9$ and $0.0975 \pm j0.88$). For the purpose of comparing how closely the bounds and estimate follow the actual peak value of the transient, the direct time-domain bound and the Fourier transform bound will be referred to as bound1 and bound2, respectively. Since the estimate and bound2 depend on the number of samples taken around the unit circle, N (which is a power of two to make use of the inverse FFT), data has been taken for different values of N . Table 1 contains the data for the fourth-order narrowband bandpass filter for the case in which the past output frequency spectrum peaks only near a pole of the reconfigured filter (similar to case 1 of Example 1). Recall that this situation

will produce a relatively well-behaved transient response. The column labeled "peak" refers to the actual peak value of the transient response. Table 2 contains data for the same filter for the case in which the past output frequency spectrum peaks away from the poles of the reconfigured filter (similar to case 2 of Example 1).

Of the two bounds and the estimate, the estimate seems to follow the actual peak value of the transient the closest for both filters under both sets of initial output states. This is true even for relatively small N . The time-domain bound, which does not depend on N , is somewhat inconsistent in its closeness to the peak value. The Fourier transform bound seems to approach the transient peak for large N but can be significantly greater than the transient peak for small N .

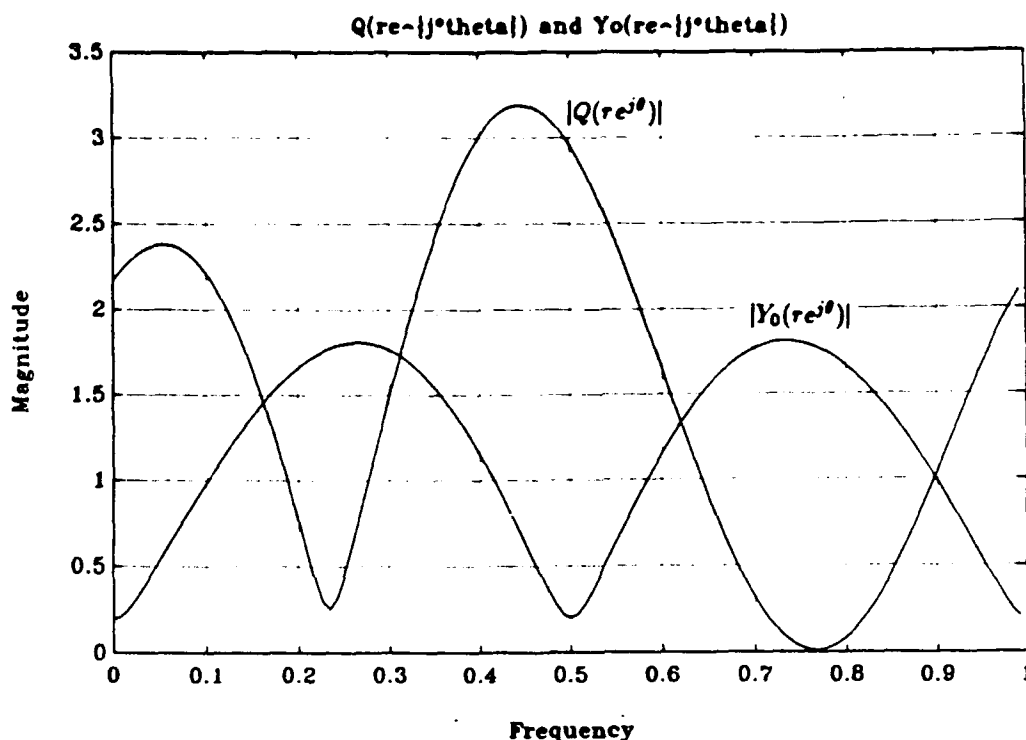


Figure 1: $|Q(re^{j\theta})|$ and $|Y_0(re^{j\theta})|$: Case 1.

5 Conclusions and Recommendations for Future Research

There is much left to be done in order to obtain a deeper understanding of the transient properties of structurally adaptive signal processing systems. The work of this effort addressed one of the simplest possible adaptations and showed that dynamic structural change was necessary to deal with the transients produced by this type of change. Clearly, this should give additional demonstration of the need for structural adaptation in dynamically changing systems.

A logical next step in the analysis of transients is to consider a linear filter in which not only the parameters but also the order (i.e., the size) of the filter changes during run-time. Actually, the majority of the analysis of this case is going to be the same as that described in this report. The main difference lies in determining what the initial state of the new filter will be. This will be a function of whether the new filter is larger or smaller than the old filter. Another research direction that is somewhat related to this one is the case where the parameters and the method of implementation of the filter change. For example, a filter implemented as a direct form II structure is changed to a different filter with a lattice structure. In this case the question of transforming the final state of the old structure into the initial state of the new structure should be investigated.

On a larger scale, the problem of redesigning entire sections of a running large-scale structurally adaptive system should be studied. It is very possible that a scheme much like the tapered switch or overlapping sum methods of this report will be applicable to this larger

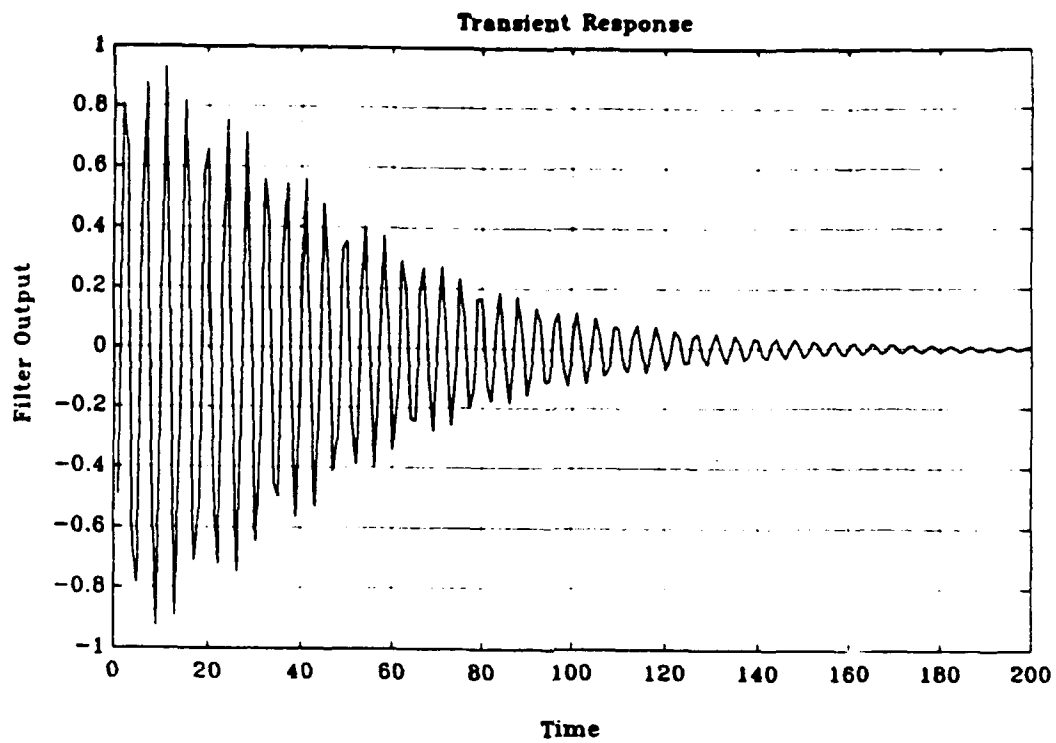


Figure 2: Transient response: Case 1.

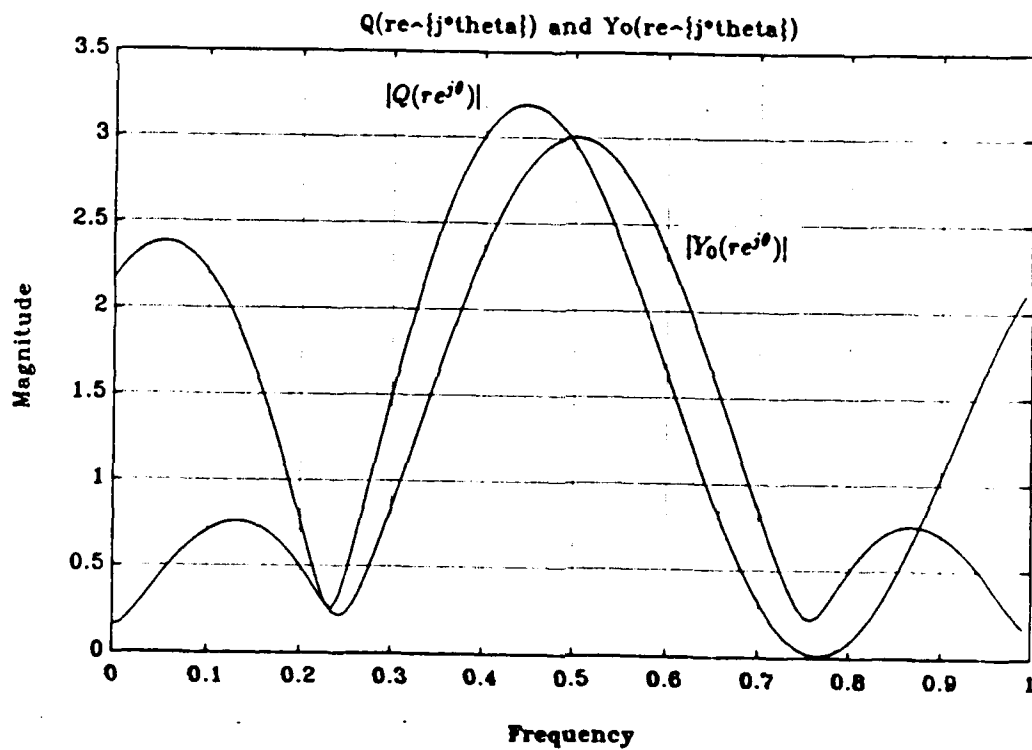


Figure 3: $|Q(re^{j\theta})|$ and $|Y_0(re^{j\theta})|$: Case 2.

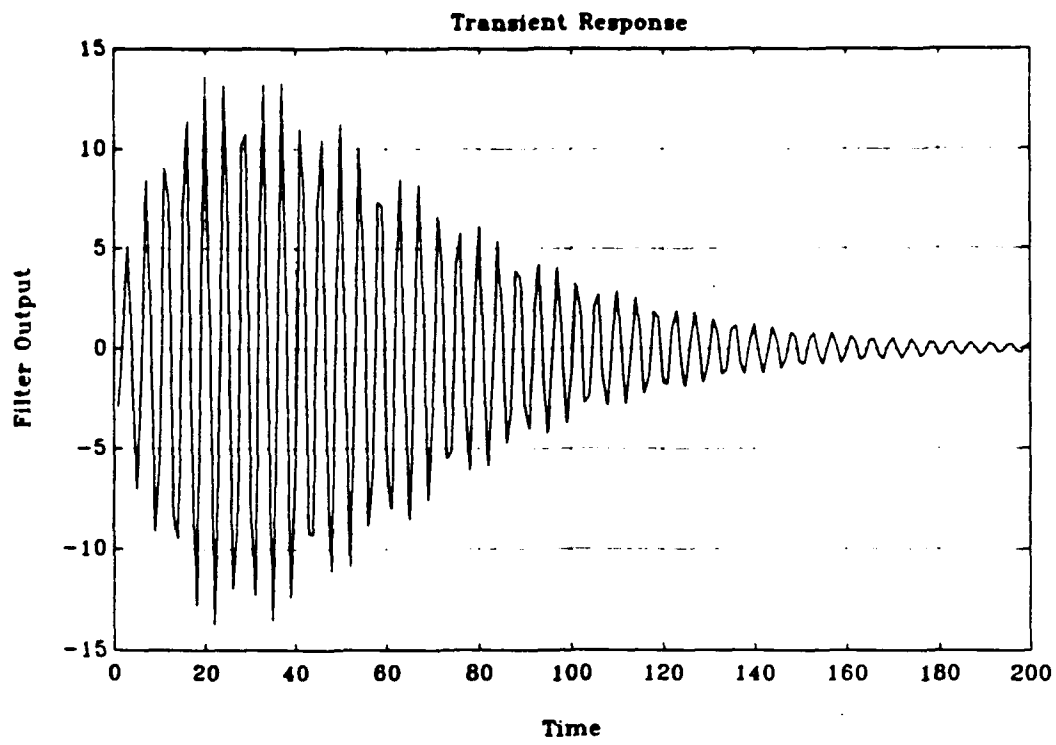


Figure 4: Transient response: Case 2:

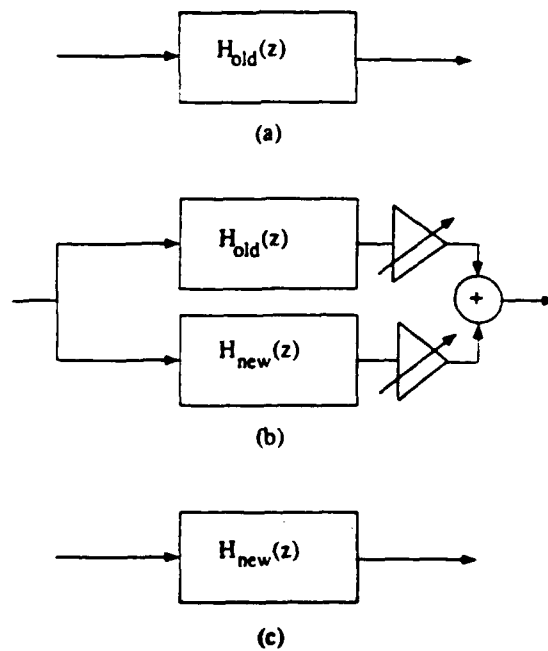


Figure 5: Block diagram for a linearly tapered switch redesign.

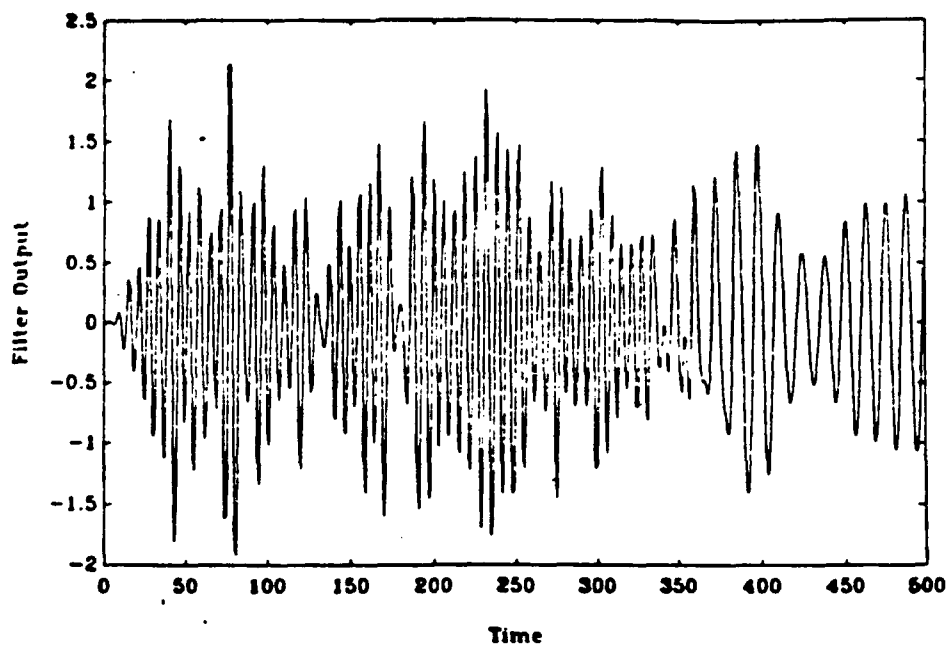


Figure 6: Filter output for tapered switch redesign.

TABLE 1
BOUND AND ESTIMATE DATA FOR FOURTH-ORDER
NARROWBAND FILTER: CASE 1.

N	Bound1	Bound2	Estimate	Peak
8	1.2228	1.6341	1.1715	0.9993
16	1.2228	1.3073	0.6483	0.9993
32	1.2228	1.3873	0.5215	0.9993
64	1.2228	3.4344	2.5015	0.9993
128	1.2228	2.5588	1.4318	0.9993
256	1.2228	2.2716	1.0919	0.9993
512	1.2228	2.1495	0.9837	0.9993
1024	1.2228	2.1419	0.99	0.9993
2048	1.2228	2.1433	0.999	0.9993

TABLE 2
BOUND AND ESTIMATE DATA FOR FOURTH-ORDER
NARROWBAND FILTER: CASE 2.

N	Bound1	Bound2	Estimate	Peak
8	62.70	8914.7	20.00	25.68
16	62.70	4455.6	20.00	25.68
32	62.70	2228.3	20.00	25.68
64	62.70	1151.3	19.70	25.68
128	62.70	558.9	20.75	25.68
256	62.70	274.8	19.73	25.68
512	62.70	135.6	20.03	25.68
1024	62.70	67.7	23.30	25.68
2048	62.70	33.9	25.70	25.68

problem. The tapered switch and overlapping sum methods are not limited to LTI systems.

6 Bibliography

References

- [1] Sztipanovits, J., and Bourne, J.R., "Architecture of Intelligent Medical Instruments," Journal of Biomedical Measurements Informatics and Control, London, UK., Vol.1, No. 3, pp. 140-146, 1987.
- [2] Biegl, C., Karsai, G., Sztipanovits, J., Bourne, J., Mushlin, R., Harrison, C., "Execution Environment for Intelligent Real-Time Systems," Proc. of the 8th Annual IEEE/EMBS Conference, Dallas, TX, pp. 807-811, 1986.
- [3] Karsai, G., Biegl, C., Sztipanovits, J., Bourne, J., Mushlin, R., Harrison, C., "Experiment Design Language for Intelligent MRI Systems," Proc. of the 8th Annual IEEE/EMBS Conference, Dallas, TX, pp. 803-807, 1986.
- [4] Sztipanovits, J., Biegl, C., Karsai, G., Bourne, J., Mushlin, R., Harrison, C., "Knowledge-Based Experiment Builder for Magnetic Resonance Imaging (MRI) Systems," Proc. of the 3rd IEEE Conference on Artificial Intelligence Applications, Orlando, FL, pp. 126-133, 1987.
- [5] J. Sztipanovits, "Toward Structural Adaptivity," Proc. of the 1988 IEEE International Symposium on Circuits and Systems, Espoo, Finland, 1988.
- [6] Sztipanovits, J., Biegl, C., Karsai, G., "Graph Model-Based Approach to Representation, Interpretation and Execution of Real-Time Signal Processing Systems", International Journal of Intelligent Systems, 1988 (in press).
- [7] W. Blokland, J. Sztipanovits, "Knowledge-Based Approach to Reconfigurable Control Systems," Proc. of the 1988 American Control Conference, Atlanta, Georgia, 1988. (in press)
- [8] J. Sztipanovits, and R.B. Purves, "Coupling Symbolic and Numeric Computations in Distributed Environment", in Kawalik, J. S., Kitzmiller, C. T., (ed.) Coupling Symbolic and Numerical Computing in Expert Systems, II, pp. 117-128, North-Holland, 1988.
- [9] J. Sztipanovits, et. al., "Programming Model for Coupled Intelligent Systems in Distributed Execution Environment," Proc. of the SPIE's Cambridge Symposium on Advances in Intelligent Robotics Systems, Cambridge, MA, 1987.

- [10] J. Sztipanovits, C. Krishnamurthy, B.R. Purves, "Testing and Validation in Artificial Intelligence Programming," Proc. of the SPIE's Cambridge Symposium on Advances in Intelligent Robotics Systems, Cambridge, MA, pp. 1987.
- [11] J. Sztipanovits, et.al., "Cooperative Systems for Real-Time Process Monitoring and Process Diagnostics", Proc. of the AI 87 JAPAN, Osaka, Japan, pp. 419-416, 1988.
- [12] G. Karsai, et.al., "Intelligent Supervisory Controller for Gas Distribution Network," Proc. of the 1987 American Control Conference, Minneapolis, Minnesota, pp. 1353-1358, 1987.
- [13] Padalkar, S., Karsai, G., Sztipanovits, J.: "Graph-Based Real-Time Fault Diagnostics," Proc. of the Fourth Conference on Artificial Intelligence for Space Applications, pp. 115-124, Huntsville, AL.
- [14] C. Biegl, "Knowledge-Based Generation of Process Simulation Models," Proc. of the 19th Southeastern Symposium on System Theory, Clemson, SC. pp.139-143, 1987.
- [15] Sztipanovits, J. and Wilkes, D. M. "Structurally Adaptive Signal Processing Systems," presented at the 1988 Digital Signal Processing Workshop, South Lake Tahoe, CA, September, 1988.
- [16] Jain, R., Goossens, G., Claesen, L., Vandewalle, J., De Man, H., Gazsi, L., and Fettweis, A., "CAD Tools for the Optimized Design of Custom VLSI Wave Digital Filters," Proc. of IEEE ICASSP 1985, pp. 1465-1468, Tampa, FL, 1985.
- [17] Péceli, G., Investigation of Recursive Signal Processing Algorithms. Technical Sciences Thesis, Budapest, 1985 (in Hungarian).
- [18] Péceli, G., "A Common Structure for Recursive Discrete Transforms," IEEE Trans. on Circuits and Systems, vol. CAS-33, no. 10, pp. 1035-1036, Oct. 1986.
- [19] Wilkes, D.M., et al., "The Multigraph and Structurally Adaptive Signal Processing," Vanderbilt University, Measurement and Computing Systems Laboratory, Technical Report # 89-004, 1989.
- [20] A. D. Koffman, A Transient Analysis of Structurally Adaptive Signal Processing Systems. MS thesis, Vanderbilt University, 1990.
- [21] Wilkes, D. M., and Koffman, A. D., "Transient Behavior of a Dynamically Reconfigured Filter," submitted to IEEE Transactions on Acoustics, Speech, and Signal Processing, in June 1990.

- [22] Proakis, J.G. and D.G. Manolakis, Introduction to Digital Signal Processing. New York: McGraw-Hill, 1988.

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Conducted by the
UNIVERSAL ENERGY SYSTEMS, INC.

GRANT REPORT

RELIABILITY IN SATELLITE COMMUNICATION NETWORKS

Prepared by:	Stephan E. Kolitz
Academic Rank:	Assistant Professor
Department and	Management Sciences Department
University:	University of Massachusetts/Boston
Date:	June 30, 1988
Contract No:	F49620-85-C-0013

RELIABILITY IN SATELLITE COMMUNICATION NETWORKS

Stephan E. Kolitz

ABSTRACT

A very important requirement of a communication network is to deliver data from a source node to a destination node. This report looks at several types of network layouts (topologies), identifies reasonable reliability measures and algorithms which calculate these measures, develops useful analytic tools, and finds a shortest-path algorithm.

I. INTRODUCTION

A very important requirement of a communication network is to deliver data from a source node to a destination node. In practice, communication networks can be categorized based on the architecture and techniques used to deliver the data. For the purpose of this report, the communication networks can be regarded as packet switching networks, but the issues that are addressed here are applicable to other types of networks. Some excellent references in this field are Rosner [1982], Tanenbaum [1981] and Stallings [1985].

The physical layout of a communication network and the method of routing data in it are clearly very closely related. Tanenbaum suggests a number of desirable attributes in the routing function of a communication network: correctness, simplicity, robustness, stability, fairness and optimality. Stallings adds efficiency to this list. There is no way to optimize routing over all these objectives; there is always a trade-off involved. However, the design of the network should allow for routing which is "good", however the "goodness" may be measured.

This report looks at several types of network layouts (topologies), identifies reasonable reliability measures and algorithms which calculate these measures, develops useful analytic tools, and finds a shortest-path algorithm.

II. COMPLEXITY THEORY

A problem which requires n bits to be fully specified is said to be of size n . If an algorithm's running time cannot be bounded by a polynomial function of n , then the algorithm is called an exponential time algorithm and is said to be intractable.

Problems that are called NP-complete form an equivalence class in the following sense: if a polynomial time algorithm exists for one of the NP-complete problems, then every NP-complete problem has a polynomial time algorithm. As of now, no polynomial time algorithm has been found for any of the known NP-complete problems. While many people believe the conjecture that NP-complete problems are intractable, it has not been proved and it is the most important open question in current complexity theory. Many network reliability problems are at least as hard to solve as the NP-complete problems.

III. GRAPHS

A graph $G = (V, E)$ consists of a vertex set

$V = \{ v(1), v(2), \dots, v(N) \}$

and an edge set

$E = \{ e(1), e(2), \dots, e(M) \}.$

Each element in E is a two-element subset of V . A directed graph (called a digraph) is a graph where E consists of ordered pairs of $v(i)$'s. A path in a graph is an alternating sequence of vertices and edges, where the vertices and edges can be labelled so that if the path is written

$v(1) \ e(1) \ v(2) \ \dots \ e(k-1) \ v(k)$ then

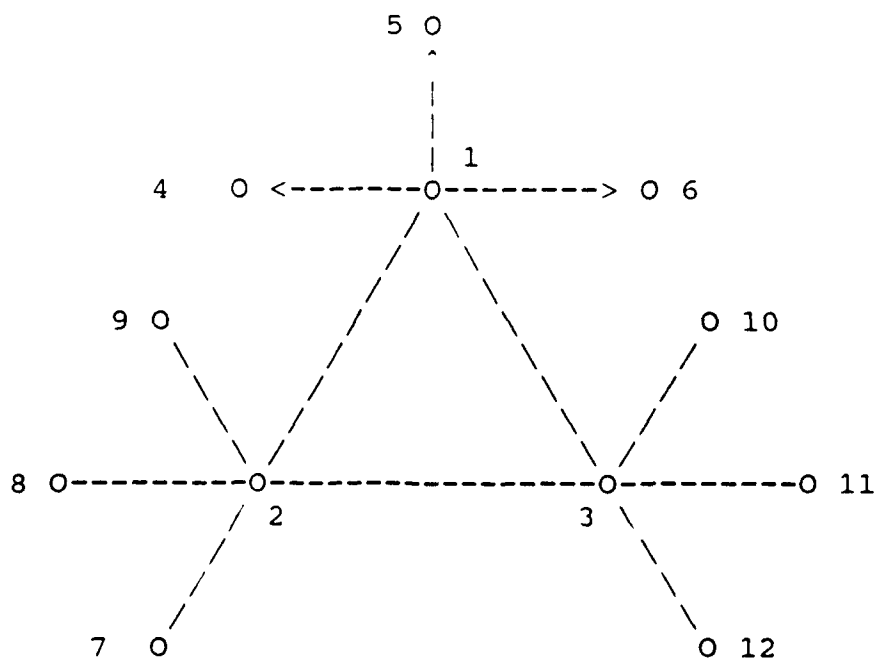
$$e(j) = \begin{cases} \{ v(j), v(j+1) \} & \text{for a graph} \\ (v(j), v(j+1)) & \text{for a digraph.} \end{cases}$$

Vertex 1 and vertex k are said to be connected if there is a path as above. A graph is connected if all possible vertex pairs are connected.

At any time of interest, the event " $v(i)$ is operative" has probability denoted $p(v(i))$ or $p(i)$; this probability is the reliability of vertex i . Similarly, $p(e(j))$ or $p(j)$, the probability that edge $e(j)$ is operative, is the reliability of edge j .

This paper is concerned with reliability problems in communication networks, hence the more suggestive terminology nodes and links will be used for the generic terms vertices and edges. In addition, the notation $v(i)$ and $e(j)$ will be simplified by suppressing the "v" and "e" unless necessary for clarity. Network will be used instead of graph, with the understanding that the network could be a digraph. The particular layout of the network will be called the network topology. A network is represented through the use of a picture in the usual way; an example is Figure 1. Note that the links from node 1 are directed while all other links are not directed; therefore a message can be sent from any node except 4, 5 and 6 to any other node if all nodes and links are operative. N, the number of nodes in the network, equals 12 in this example.

FIGURE 1. An example of a network.



One example of a network reliability problem is as follows.

Let G be a network with known node and link reliabilities. What is the probability that the network is connected?

This problem and most other network reliability problems are at least as hard as the NP-complete problems. Hence the existence of an algorithm which solves this problem in polynomial time would imply the existence of polynomial time algorithms for all NP-complete problems.

This is the fundamental problem with virtually all network reliability problems; no one has found polynomial time algorithms which solve them, and furthermore most researchers feel that these problems are intractable.

IV. A TRACTABLE NETWORK RELIABILITY PROBLEM

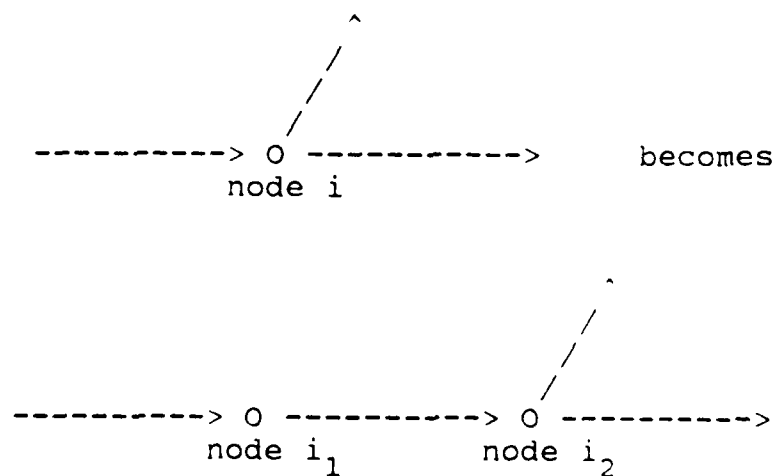
The following is given:

- 1) the network topology
- 2) the reliability of all nodes and links.

The objective is to route messages by the most reliable path.

Any network with node and link failures are can be transformed into a network with only link failures. Replace every node i in the original network with two nodes i_1 and i_2 ; the reliability of the original node is assigned to the link connecting the nodes in the new network. Thus, if $p(v(i)) = p$ in the original network then in the new network the probability of the link between i_1 and i_2 being operative is $p((v(i_1), v(i_2))) = p$. An example follows.

FIGURE 2. Replacement of node reliability by link reliability



Thus, without loss of generality, we can make the very useful assumption that any network has only link failures. Suppose there exists such a network. Then

let $W(s,t) = \{ w(s,t) \mid w(s,t) \text{ is a path from node } s \text{ to node } t \}$.

Let $R(s,t) = \max \{ \begin{array}{c} p(e(j)) \\ \text{all } e(j) \text{ in} \\ w(s,t) \text{ in } W(s,t) \end{array} \}$.

$R(s,t)$ is the reliability of the most reliable path from node s to node t ; i.e. the path with the largest product link probabilities (a subset of which were node probabilities in the original network). The most reliable path is denoted $w^*(s,t)$.

Let $d(e(j)) = -\ln(p(e(j)))$ and $D(s,t) = -\ln(R(s,t))$.

Then

$D(s,t) = \min \{ \begin{array}{c} d(e(j)) \\ \text{all } e(j) \text{ in} \\ w(s,t) \text{ in } W(s,t) \end{array} \}$

is the "shortest" route from s to t in the network with distances given by the function d above.

This is the well known shortest route problem and can be solved very efficiently with existing algorithms.

The algorithm presented below is a modification of Dijkstra's label setting algorithm for finding the shortest route between nodes in a network. For large sparse matrices, the Bellman-Moore label correcting algorithm as improved by d'Esopo and coded by

Pape [1980] appears to be faster. Recent work by Glover [1984] indicates that THRESH, a hybrid of label setting and label correcting algorithms, is the currently fastest available algorithm in general. An excellent reference for this area is Syslo [1983]; it includes all but the latest work by Glover.

Assume now that there exists a directed network with only node failures. While this assumption is made primarily for ease of exposition, it is not necessarily a bad assumption for a model of a communications network based in space. The probability of node i being operative is denoted $p(i)$. The nodes are labelled $1, 2, \dots, N$.

ALGORITHM

Let $u(i, j) = \begin{cases} 0 & \text{if there is no link from } i \text{ to } j \\ p(j) & \text{if there is a link from } i \text{ to } j \end{cases}$

and $T = N$ (the set of nodes).

1) Set $r(s) = 1$ and $r(j) = u(s, j)$ for $j > 1$ where j is an element of $T = N - \{s\}$. The set T is labelled the set of temporary nodes.

2) Find i in T such that

$$r(i) = \max \{ r(j) \mid j \text{ is in } T \}.$$

3) Set $T \leftarrow T - \{i\}$. If T is the empty set, then stop;
else go to step 4. Node i is labelled permanent.

4) For each j in T , set

$$r(j) \leftarrow \max \{ r(j), r(i)u(i,j) \}$$

5) Go to Step 2.

Optimal routes are generated by recording the nodes which solve the maximization problem in Step 4) above. The time complexity of this algorithm is $O(N^3)$, which includes finding the most reliable path from every node to every other node.

V. RELIABILITY MEASURES AND ALGORITHMS

Network reliability analysis is very difficult for a number of reasons. First of all, it is not easy to even define what the problem is. Secondly, all of the measures presented below result in problems that are at least as hard as the NP-complete problems. In addition, solution algorithms do not lend themselves to optimization of network design, but rather to a description of proposed topologies. In practice, a combination of analysis and simulation is used to try to produce one network design. In Sections VII and VIII, some useful analytic tools are developed.

There is a large growing literature in network reliability. One or more of the following reliability measures have appeared in many recent papers. (See Bibliography section on reliability.)

- 1) the probability that all nodes are communicating
- 2) the probability that all operative nodes are communicating
- 3) the probability that all operative nodes are communicating with a given node
- 4) the expected number of nodes communicating
- 5) the expected number of nodes communicating with a given node
- 6) the expected number of node pairs communicating with a given node
- 7) the probability that the system operates

Two of the best papers are Ball [1979] and Ball and Nemhauser [1979]. Ball [1979] specialized algorithms which calculate the first six measures for networks in which only nodes can fail. The foundation of these and virtually all other reliability algorithms is clever partitioning of the sample space and appropriate conditioning of events.

The measure addressed here is 1), the probability that all nodes are communicating. With only link failures, this is also the probability that the network is connected.

VI. TOPOLOGIES

Loop topologies have received a fair amount of attention in recent years. There are a series of papers which deal with loop topologies set up without a central control node. This enables routing schemes to be set up which are adaptive to changes (additions or deletions) in the network. Loop topologies can be used in networks ranging from local area networks to world-wide satellite communication networks.

Saltzer [1981] found some good reasons for using a ring (or loop) network in local area networks. Raghavendra [1981] proved that the optimal loop topology should have forward short links to the next node and backward long links to the node $\lfloor \sqrt{N} \rfloor$ nodes away. Optimal here means that this topology minimizes the maximum distance between nodes. Claims are made for optimality in terms of reliability, but the reliability measure used is very primitive.

Brayer [1984] took this result and a routing algorithm from Chyung [1975] to develop an algorithm which automatically updates routing when there are either additions to or deletions from the network. Raghavendra [1985] studied the performance of the double loop network topology using a variety of adaptive routing algorithms. He compared utilization and delay figures derived from a simple queueing model to estimated values based on a series of simulations.

In what follows, loop topologies follow the optimal layout, but with no directed links. (If the nodes are drawn in a circle and connected as below, it is clear why the term loop topology is used.) A city topology is one which resembles the classic street structure of a number of cities. Note also that a city topology can be regarded as a subset of a loop topology, where the boundary nodes do not have the full connection of a loop topology. The distance between two nodes along a given path is measured by the number of links in the path, in both topologies. The topologies do not necessarily have to be square.

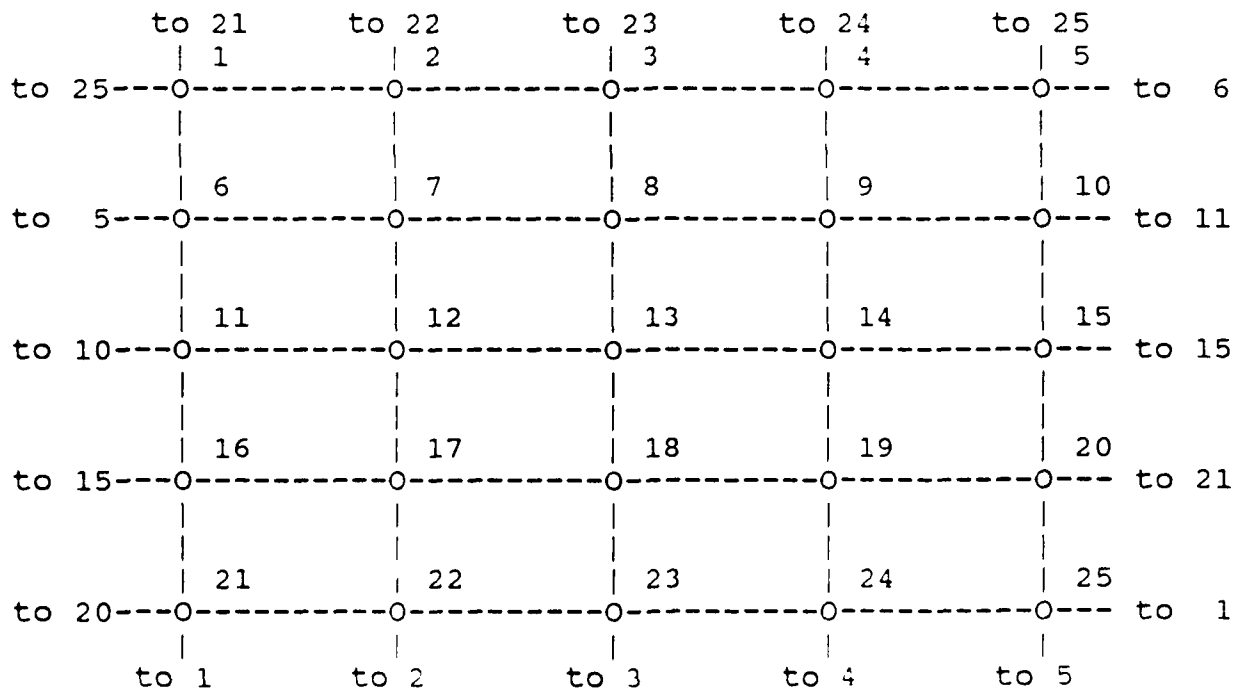
City Topology

The graph below is described as a 5 by 5 city topology. Rows and columns refer to the obvious: the first row contains nodes 1 through 5, the second 6 through 10 and so on; the first column is 1,6,11,16,21.



Loop Topology

The graph below is described as a 5 by 5 loop topology.



VII. AN ANALYTIC APPROXIMATION

There are N nodes in a connected network; at node i there are $m(i)$ outbound links and link (i,j) has reliability $p(i,j)$. Then the network reliability, the probability that the network is connected (P_c) is approximately:

$$\prod_{i=1}^N \left(1 - \prod_{j=1}^{m(i)} (1 - p(i,j)) \right)$$

This approximation is based on defining non-uniform Bernoulli trials at every node, where a trial is "observing an outbound link" and a success is "the link working."

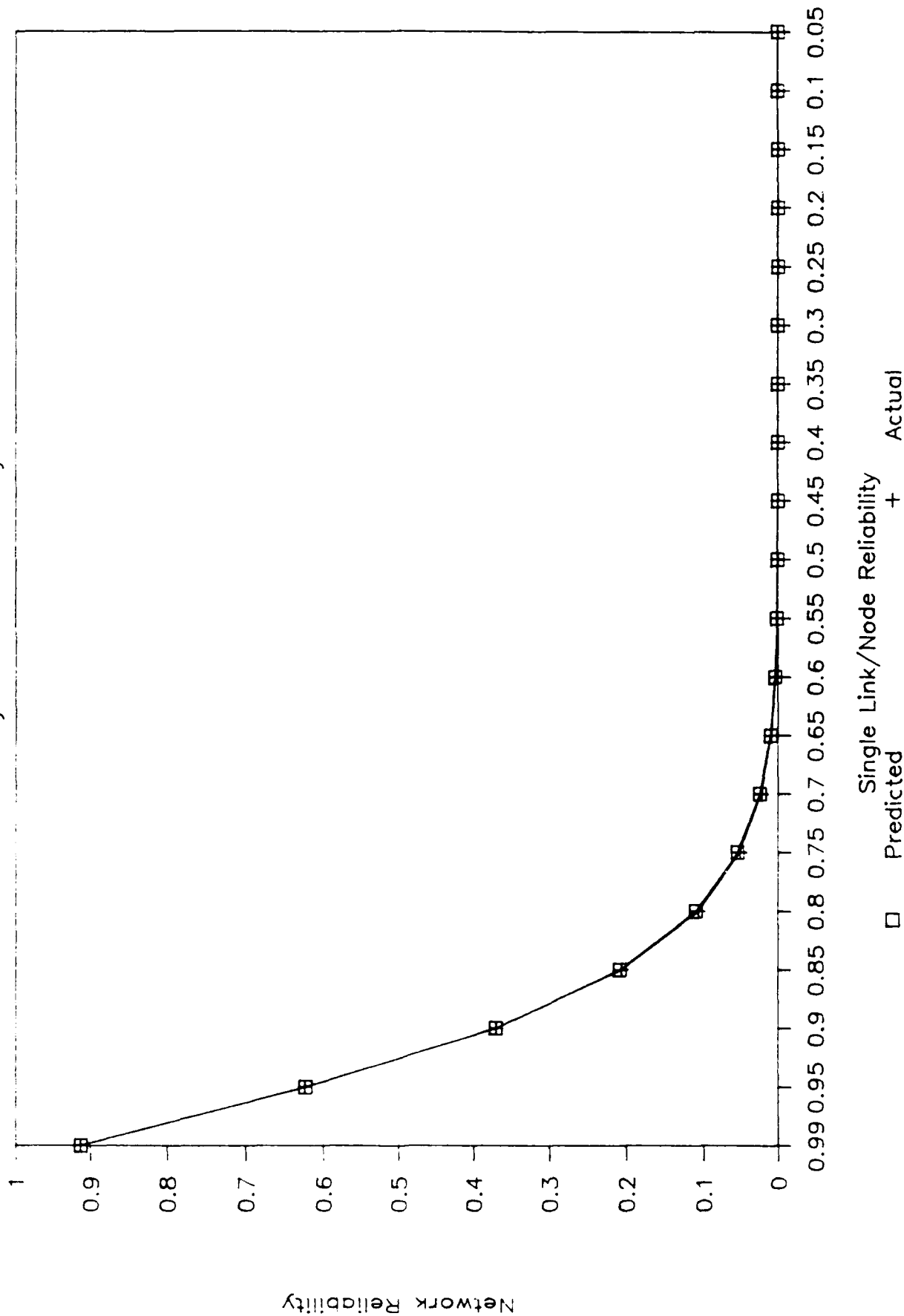
This approximation was compared to actual values calculated using an algorithm in Ball [1979]. The code for the algorithm was graciously supplied by Dr. Ball. Note that the probability is reliability measure 1) of Section V.

Some typical results are illustrated by the following graphs. The vertical axis is P_c and the horizontal axis is the single link (and node, when applicable) reliability, assumed to be the same for all links (and nodes, when applicable). A number of loop and city topologies were considered. The legends "Predicted" and "Actual" refer to the approximation and the Ball

algorithm respectively. Going beyond the size 5 by 5 was impractical for the Ball algorithm; this is due to the problem complexity discussed in Section V. The approximation does quite well for this class of topologies, and even when the approximation is at its worst, it still finds the relatively flat part of each P_c curve where reliability values are close to one. This is useful when designing such networks, as it provides a range of link and node reliabilities that result in essentially equivalent network reliability.

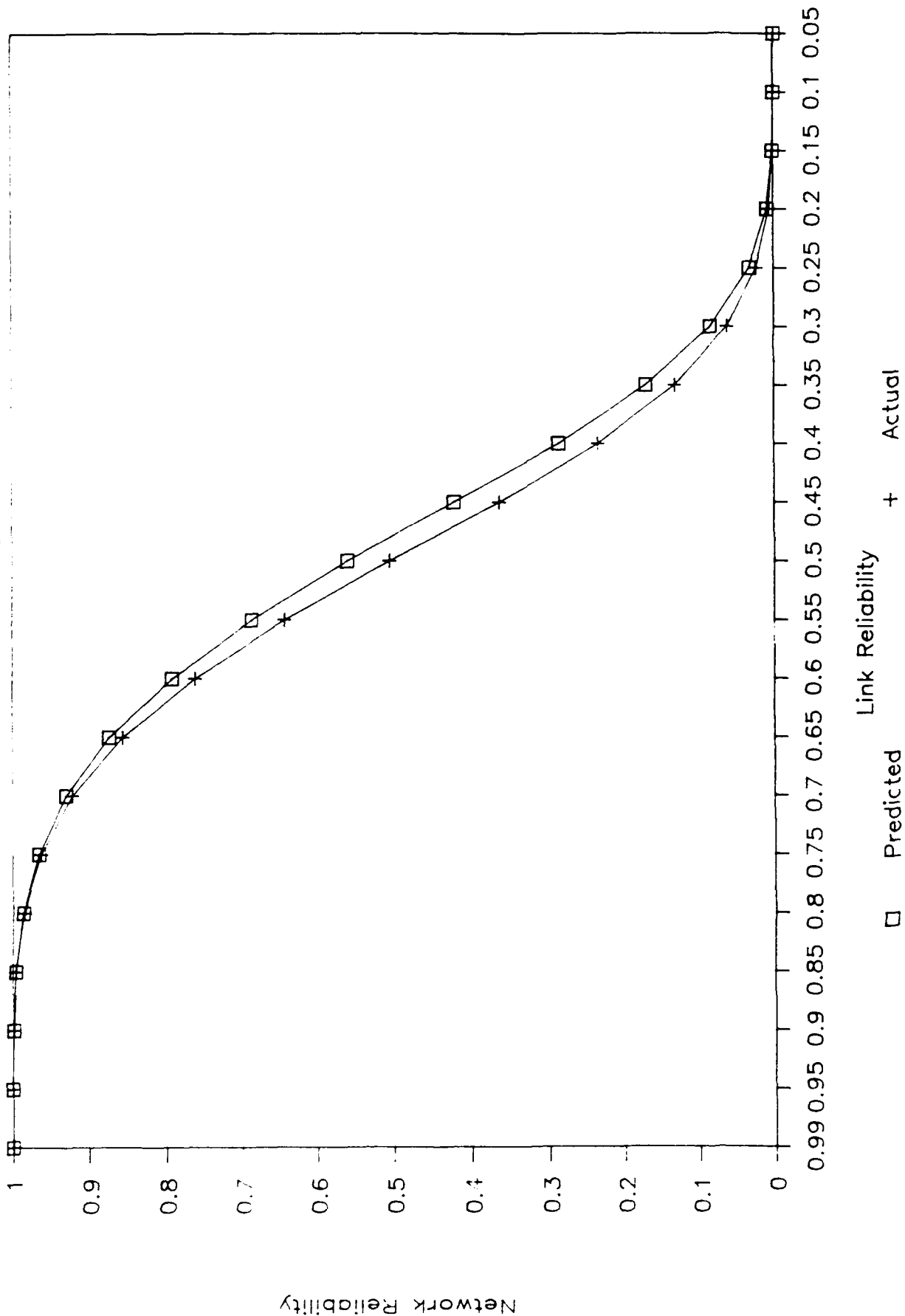
Network Reliability

3x3 City With Node Reliability



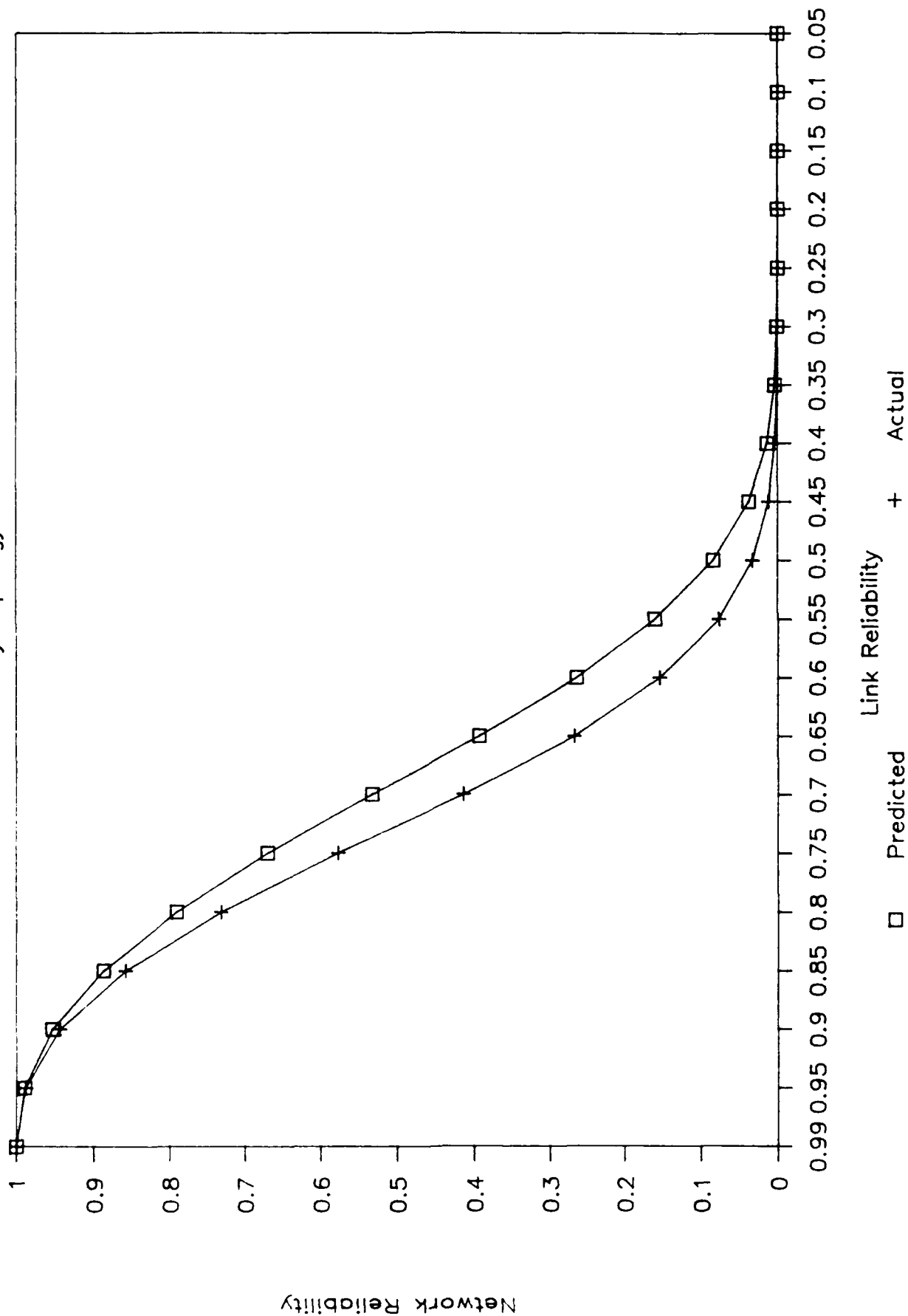
Network Reliability

3x3 Loop Topology



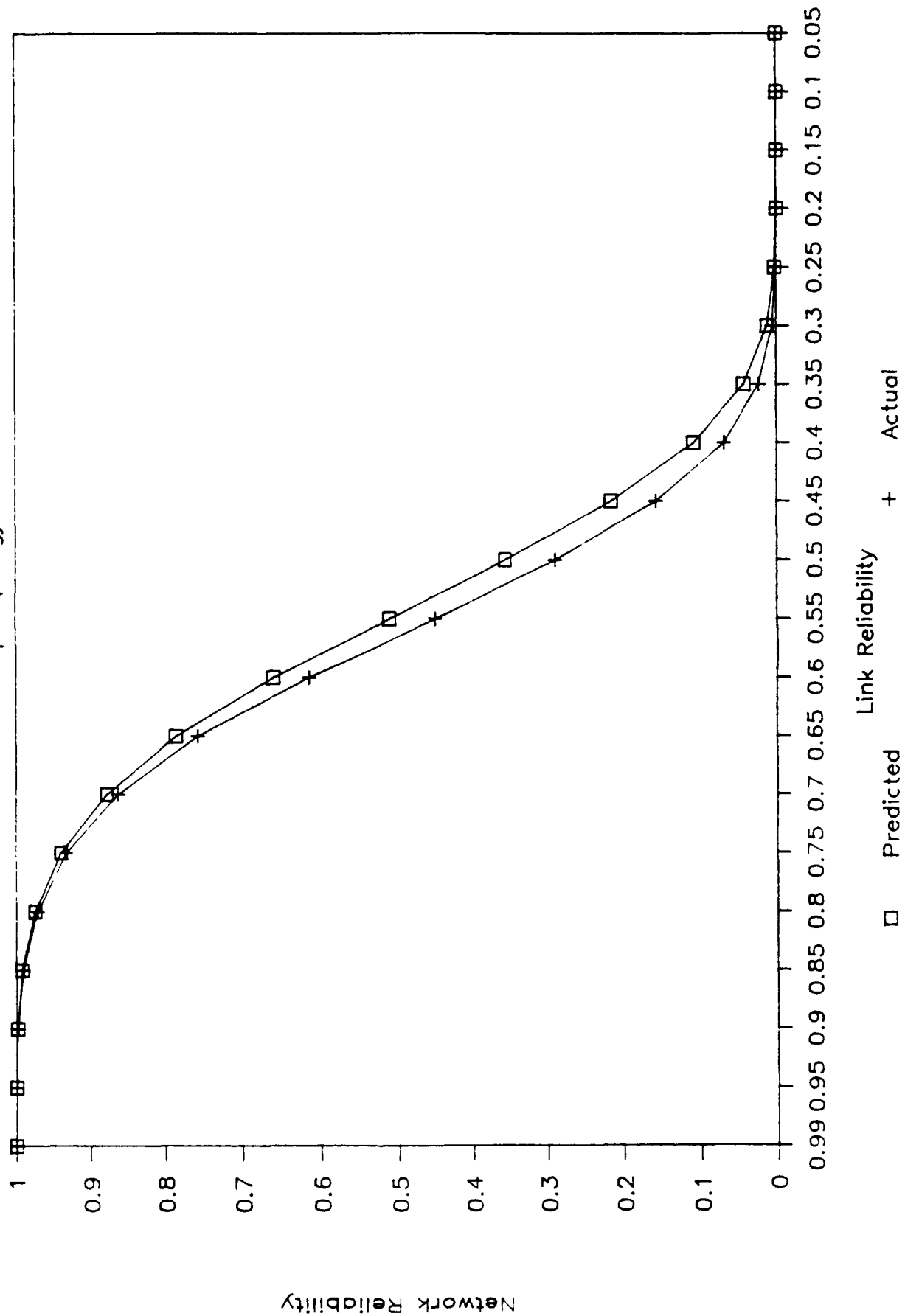
Network Reliability

4x4 City Topology



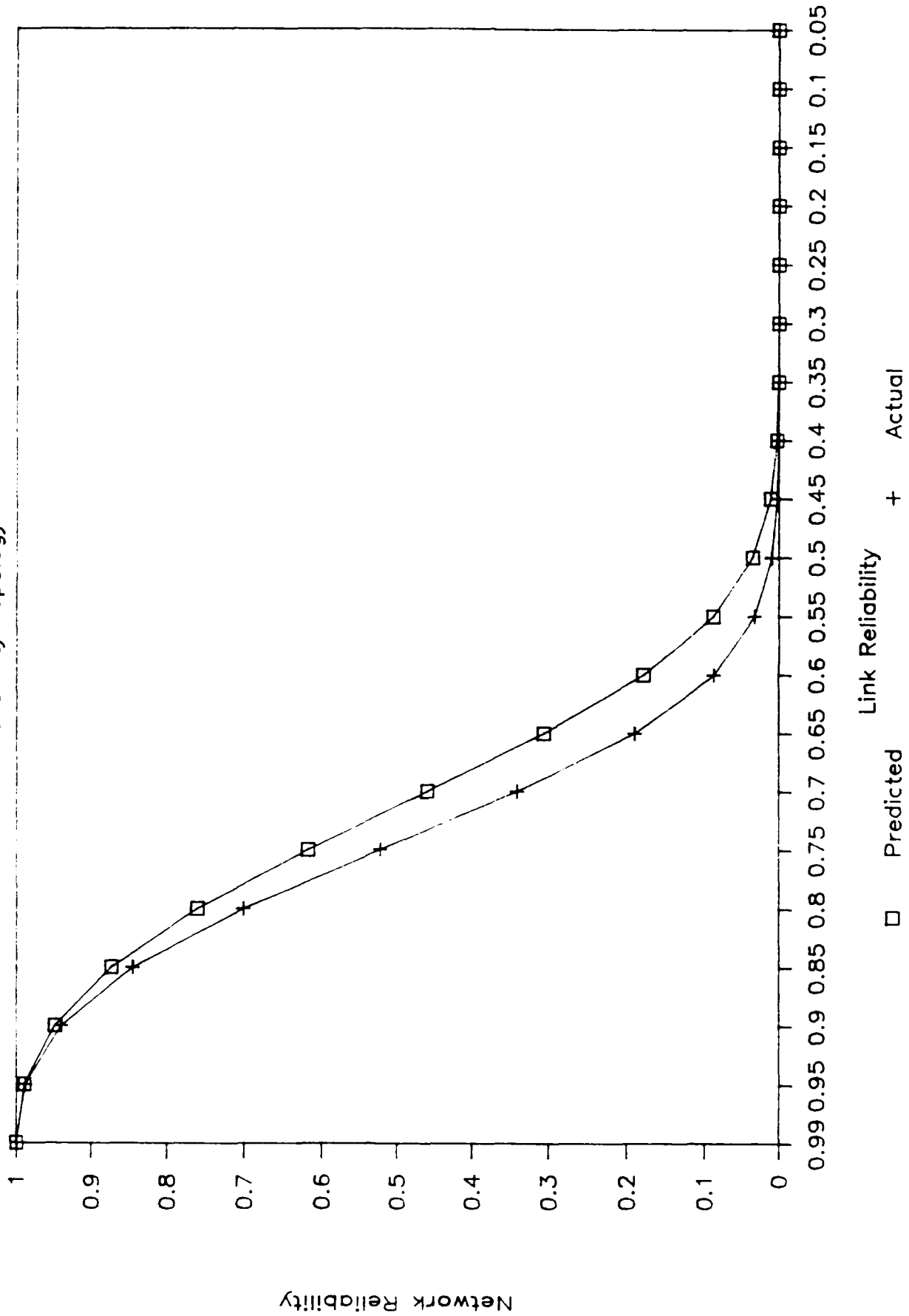
Network Reliability

4x4 Loop Topology



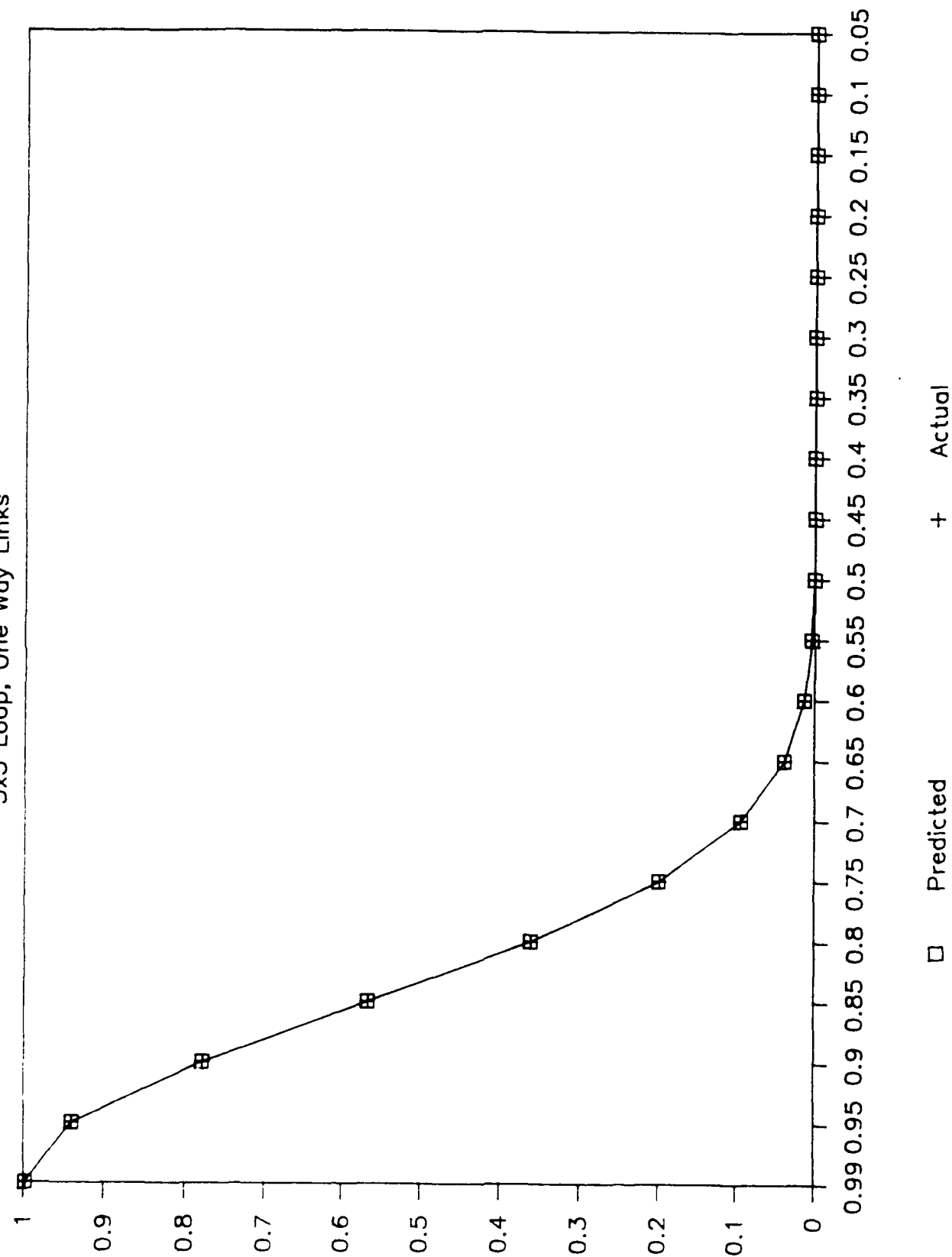
Network Reliability

5x5 City Topology



Network Reliability

5x5 Loop; One Way Links



VIII. NETWORK DESIGN

One network design problem can be stated as follows.

Given a limit to the number of links which can come out of any node, what is the topology which maximizes the probability that the network is connected?

This problem is formulated as PROBLEM I.

PROBLEM I

Let

i and j be nodes, i not the same as j throughout

$$x(i,j) = \begin{cases} 1 & \text{if there is a link from i to j} \\ 0 & \text{if there is no link} \end{cases}$$

$$p(i,j) = \text{probability that link i,j is up}$$

Then

$$\max \quad \prod_{i=1}^N \left(1 - \prod_{j=1}^N (1 - p(i,j))^{x(i,j)} \right)$$

$$\text{subject to } \sum_{j=1}^N x(i,j) \leq a(i)$$

$$x(i,j) = 0 \text{ or } 1$$

The objective function is separable, which allows for easy analysis. Also, a number of the results from Kolitz [1987] can be applied to this problem; conclusions follow.

- 1) Use a maximum marginal return algorithm, essentially Algorithm I in Kolitz [1987], to optimally assign links.
- 2) If $p(i,j)$ is constant for all i and j , then the number of links from any node should differ by no more than one from the number of links from any other node.
- 3) The optimal design equalizes reliability at each node; that is, the probability of at least one link being up at each node is the same.

Another network design problem can be stated as follows.

Given a minimum necessary level of reliability for a network, what is the topology which minimizes the cost of building the network?

This problem is formulated as PROBLEM II. Problem II is currently being studied.

PROBLEM II

Let

i and j be nodes, i not the same as j throughout

$$x(i,j) = \begin{cases} 1 & \text{if there is a link from } i \text{ to } j \\ 0 & \text{if there is no link} \end{cases}$$

$p(i,j)$ = probability that link i,j is up

$c(i,j)$ = cost of link i,j

R = the required level of network reliability

Then

$$\min \sum_{i=1}^N \sum_{j=1}^N c(i,j)x(i,j)$$

subject to

$$\sum_{i=1}^N \sum_{j=1}^N (1 - p(i,j))^{x(i,j)} \geq R$$

$$x(i,j) = 0 \text{ or } 1$$

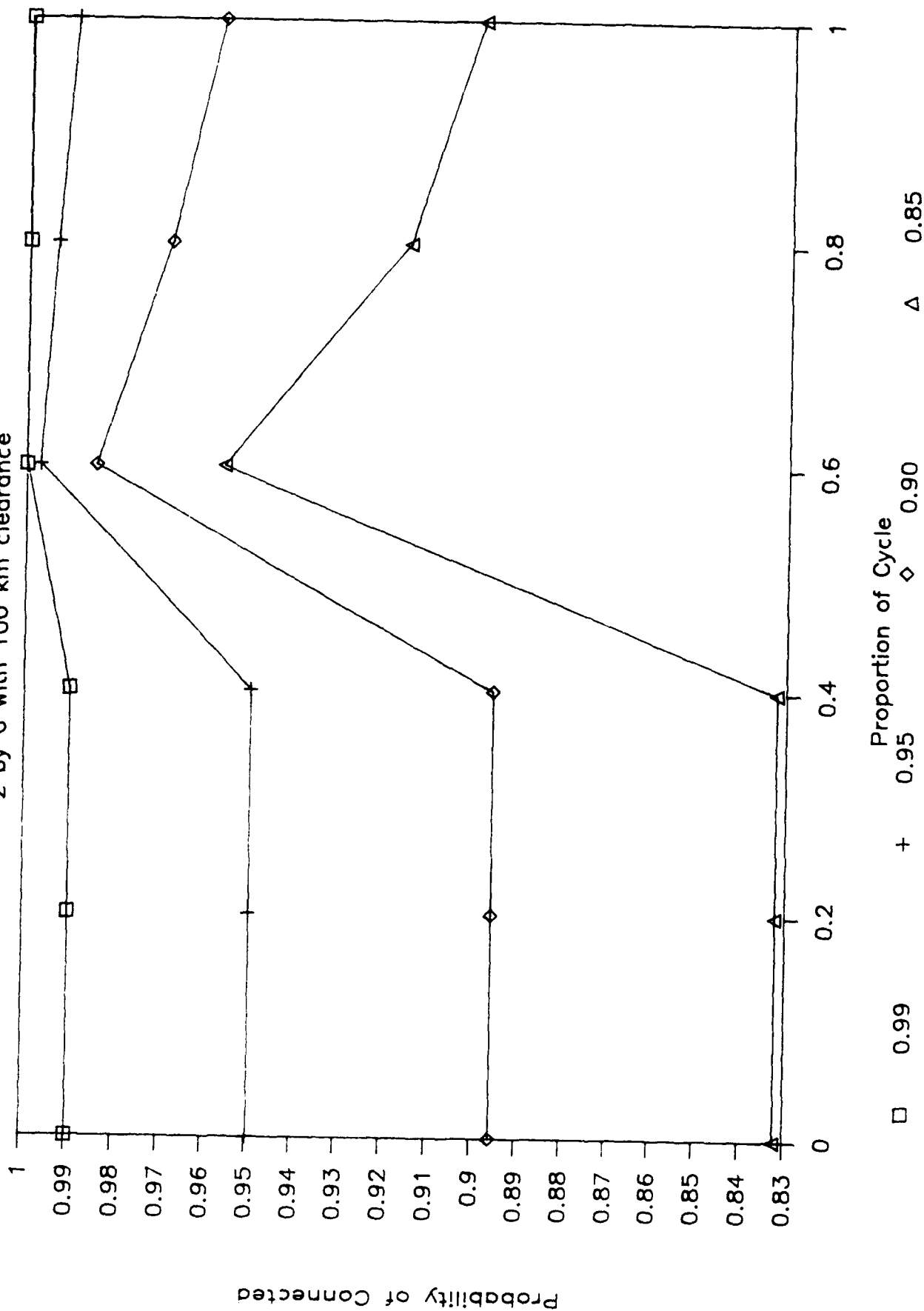
IX. SATELLITE NETWORK ARCHITECTURES

Network reliability (P_c) was calculated for two satellite network architectures. In both architectures, the orbital height is 2100 kilometers, the satellites are in polar orbit and there are two orbital planes (rings). Also, spacing is such that there is a constant time between satellites passing over the North Pole. One architecture has six satellites per ring and is denoted 2 by 6 or 2x6; the other has nine satellites per ring and is denoted 2 by 9 or 2x9. Network topology is determined by the geometry. Links are present whenever two satellites are in line of sight, subject to the line being a given minimum distance above the earth's surface. Three minimum distances are used: 0, 100 and 700 kilometers.

Because of the geometry, patterns repeat in cycles throughout a complete orbit of any one satellite. The horizontal axis is time, expressed as the proportion of a cycle; a complete orbit's graph would consist of repetitions of the graphs presented. For display purposes, the points are connected by straight lines to emphasize the changes in P_c as a function of time. The legends 0.99, 0.95, etc. are the single link reliability values.

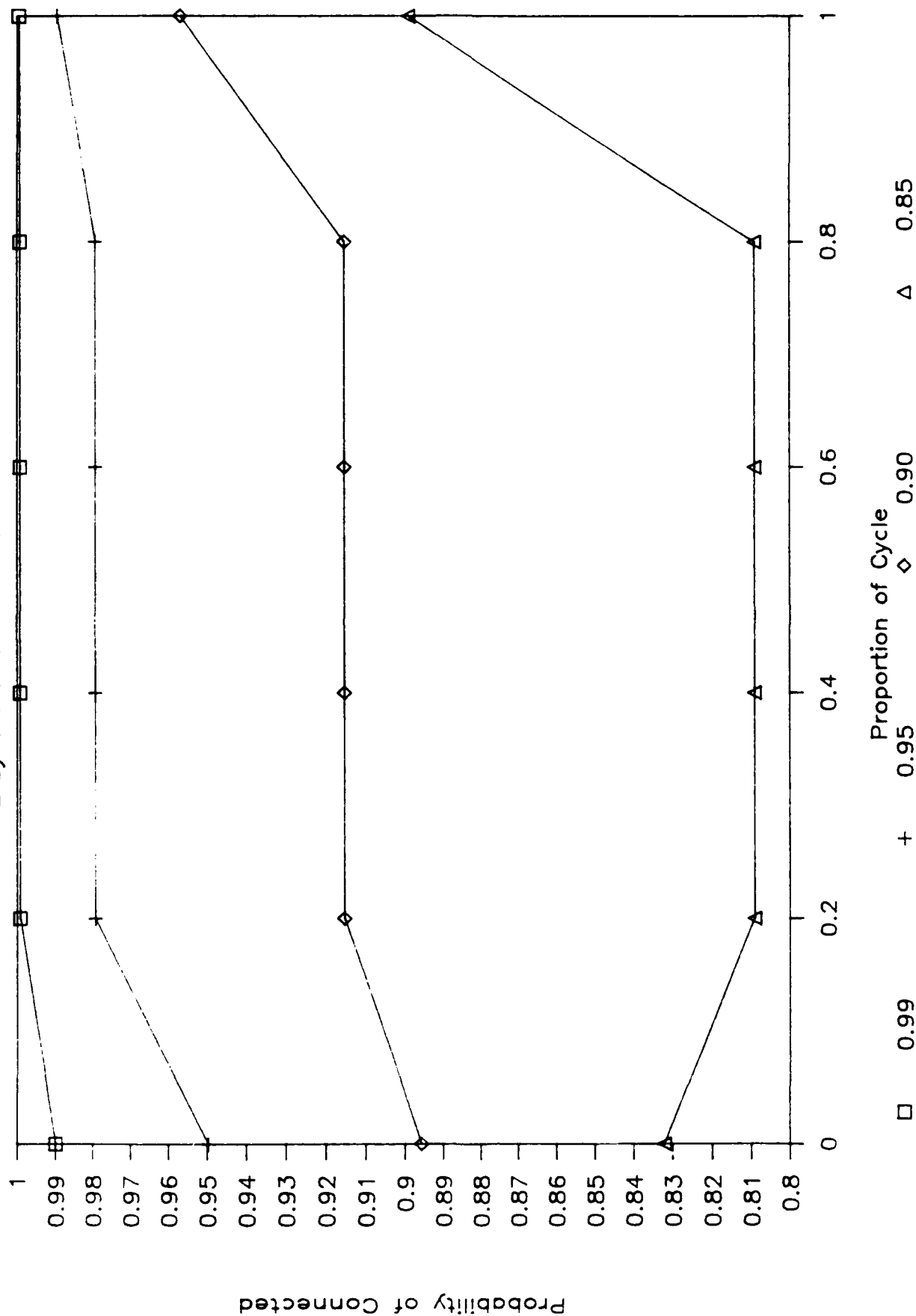
Network Reliability

2 by 6 with 100 km clearance



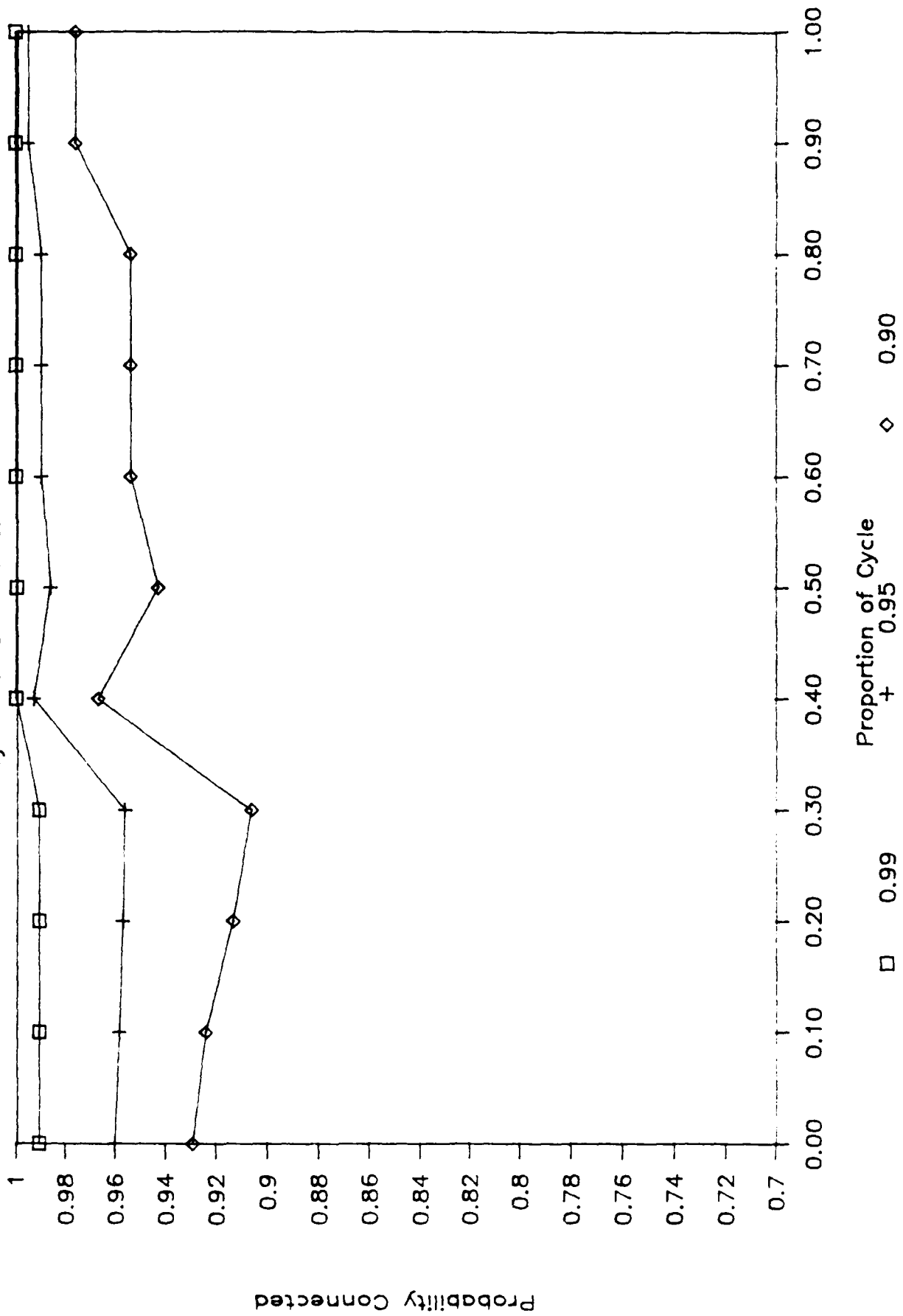
Network Reliability

2 by 6 with 700 km clearance



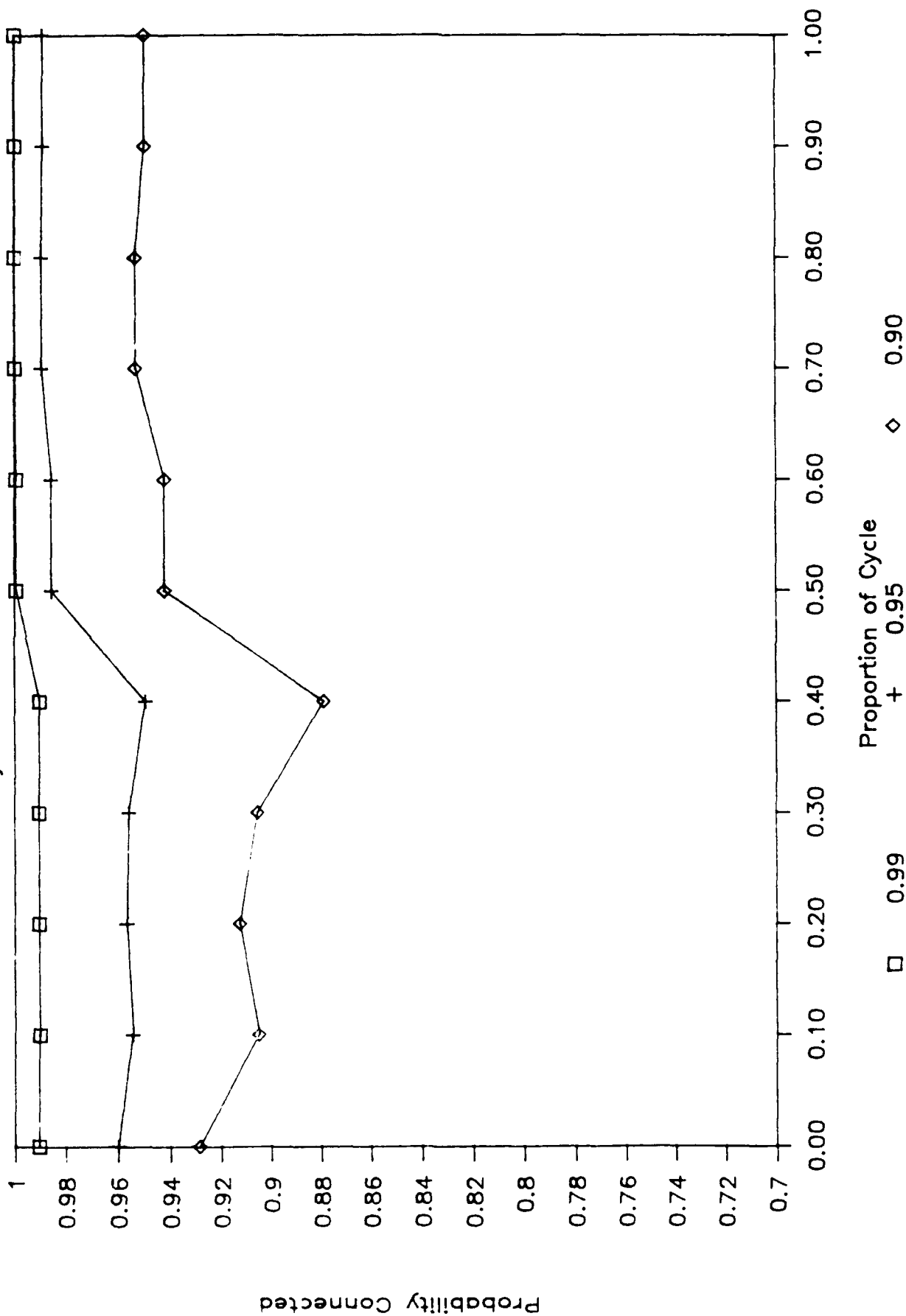
Network Reliability

2 by 9 with 0 km clearance



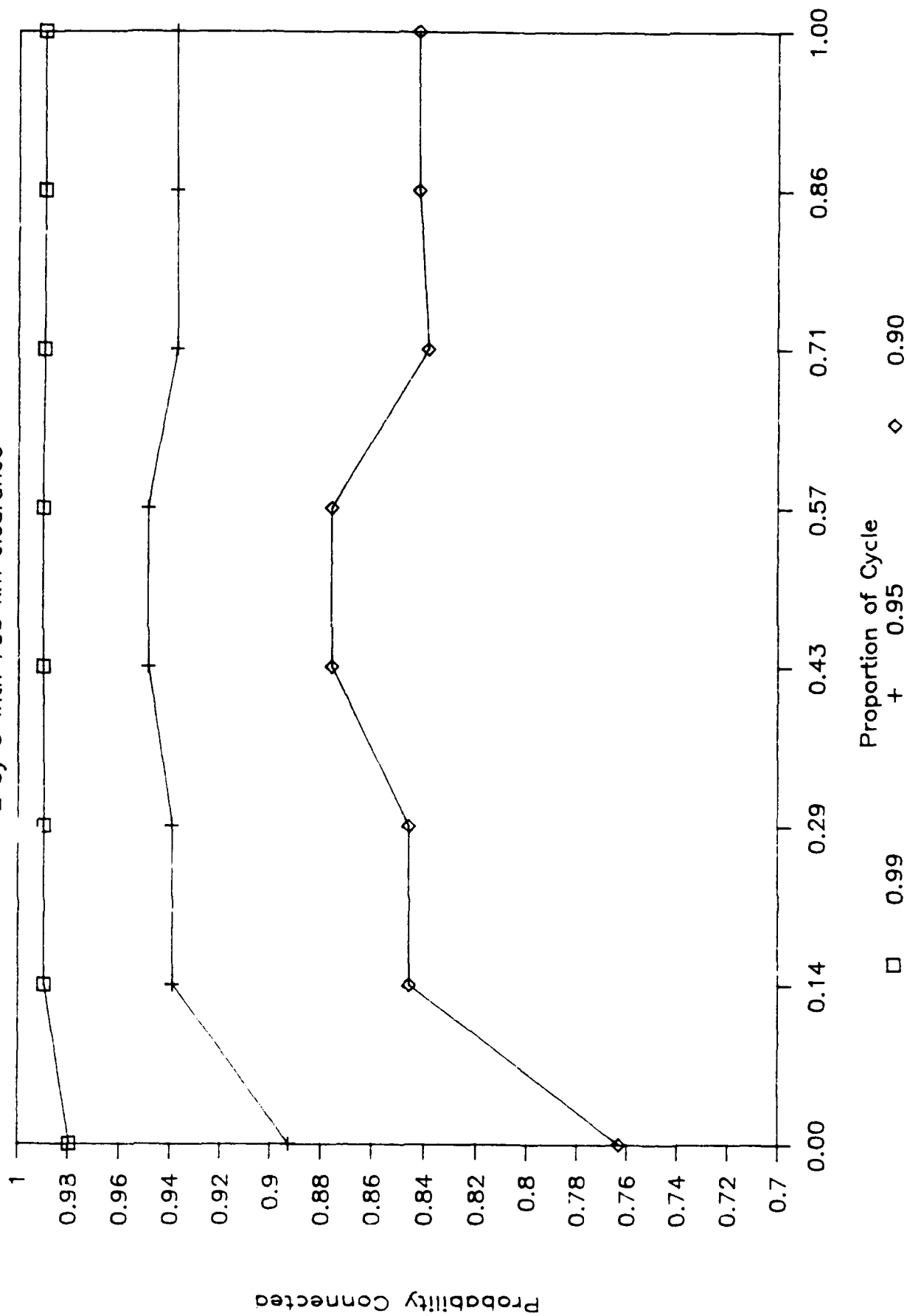
Network Reliability

2 by 9 with 100 km clearance



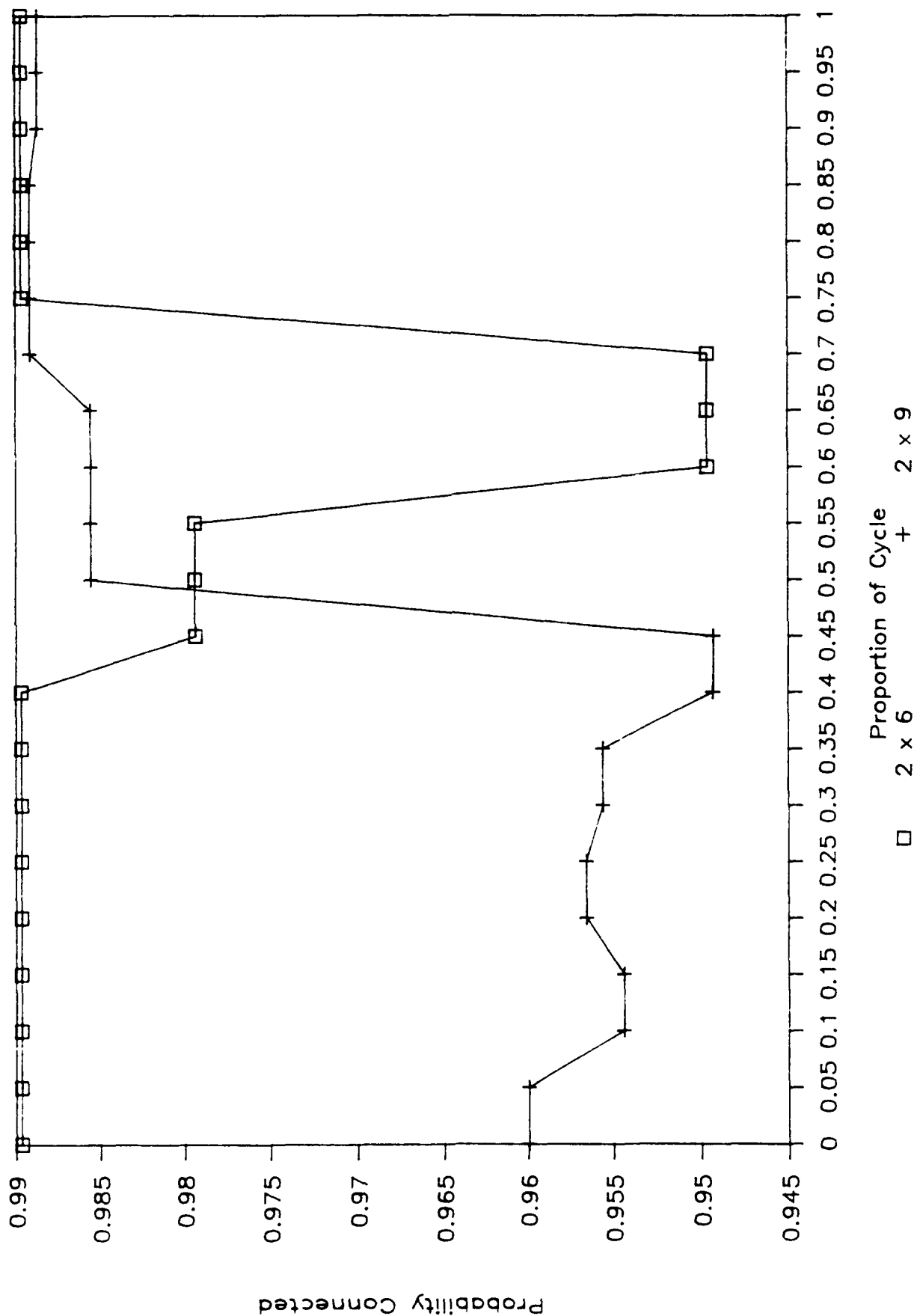
Network Reliability

2 by 9 with 700 km clearance



Network Reliability at 0.95

2x6 and 2x9 with 100 km clearance



X. ROUTING

The following algorithm is useful for making routing decisions when links and nodes are subject to failure and all link distances are one. It is very efficient when the network has the city topology

ALGORITHM

- I. Find all shortest length paths from s to t for all nodes s and t.

In the case of the the city topology, these paths are trivial to compute.

- II. Keep track of how many paths there are for each pair, and index each path by node and link.

In the case of the city topology, the number of shortest paths can be computed at the start using the following formula:

number of shortest paths from i to j = number of
combinations of A things taken B at a time where

A = distance between i and j

$$= |\text{row}(i) - \text{row}(j)| + |\text{col}(i) - \text{col}(j)|$$

$$B = \text{minimum} (|\text{row}(i) - \text{row}(j)| , |\text{col}(i) - \text{col}(j)|)$$

III. Let $N(t) = \{i: d(i,t) = 1\}$; $N(t)$ is the set of nodes one link away from t .

When a node or link goes down, delete all paths which contain it and update the number of shortest length paths for every node pair. If no shortest length paths are left, then form a new set of shortest length paths by concatenating the destination node to the shortest length paths from s to $N(t)$.

Go to II.

The advantage of this algorithm over the standard algorithms, which are very fast for the special structure here, is that all feasible shortest paths are always known. Thus, routing can be done using other criteria, such as equalization of traffic density, over different shortest paths. The problem is changed

from one of computation, using the classical methods, to one of accessing a (potentially large) database. The algorithm was coded and tested on city topologies, with good results.

BIBLIOGRAPHY

Kolitz, Stephan E., "The Multi-Weapon Multi-Target Multi-Phase Assignment Problem," AFOSR/UES Grant Report, (1987)

DATA COMMUNICATIONS

Rosner, Roy D., Packet Switching, Lifetime Learning Publications, Belmont, California, 1982

Stallings, William, Data and Computer Communications, Macmillan Publishing Company, New York, New York, 1985

Tanenbaum, Andrew S., Computer Networks, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1981

LOOP TOPOLOGIES

Brayer, Kenneth, "Packet Switching for Mobile Earth Stations Via Low-Orbit Satellite Network," Proceedings of the IEEE, Vol. 72, No. 11, pp. 1627-1636 (1984)

Brayer, Kenneth, "Autonomous Adaptive Local Area Networking: Ring Communications Via Point-To-Point Implementation," Proc. INFOCOM, pp. 49-58, (April 1984)

Chyung, Dong H. and Sudhakar M. Reddy, "A Routing Algorithm for Computer Communication Networks," IEEE Transactions on Communications, Vol COM-23, pp. 1371-1373 (1975)

Raghavendra, C.S. and M. Gerla, "Optimal Loop Topologies for Distributed Systems," ACM SIGCOMM, Vol. 11, No. 4, pp. 218-223 (1981)

Raghavendra, C.S. and J.A. Silvester, "Double Loop Network Architectures-- A Performance Study," IEEE Transactions on Communications, Vol COM-33, pp. 185-187 (1985)

Silvester, J.A. and C.S. Raghavendra, "Analysis and Simulation of a Class of Double Loop Network Architectures", Proc. INFOCOM, pp. 30-35, (April 1984)

Saltzer, Jerome H. and David D. Clark, "Why a Ring?", ACM SIGCOMM, Vol. 11, pp. 211-217, (1981)

COMPLEXITY THEORY

Garey, Michael R. and David S. Johnson, Computers and Intractability, W.H. Freeman and Company, 1979

Even, S., O. Goldreich, S. Moran and P. Tong, "On the NP-Completeness of Certain Network Testing Problems," Networks, Vol. 14, PP. 1-24 (1984)

RELIABILITY

Aggarwal, K.K. and K.B. Misra, and J.S. Gupta, "A Fast Algorithm for Reliability Evaluation," IEEE Transactions on Reliability, Vol. R-24, pp. 83-85 (1975)

Agrawal, Avinash and Richard E. Barlow, "A Survey of Network Reliability and Domination Theory", Operations Research, Vol. 32, pp. 478-492, (1984)

Agrawal, Avinash and A. Satyanarayana, "An $O(|E|)$ Time Algorithm for Computing the Reliability of a Class of Directed Networks," Operations Research, Vol 32, pp. 493-515, (1984)

Ball, Michael O., "Computing Network Reliability," Operations Research, Vol. 27, pp. 323-338, (1979)

Ball, Michael O. and George L. Nemhauser, "Matroids and a Reliability Analysis Problem," Mathematics of Operations Research, Vol 4, pp. 132-143 (1979)

Ball, Michael O., "Complexity of Network Reliability Computations," Networks, Vol. 10, pp. 153-165, (1980)

Boesch, Frank T., Frank Harary and Jerald A. Kabell, "Graphs as Models of Communication Network Vulnerability: Connectivity and Persistence," Networks, Vol. 11, pp. 57-63, (1981)

Buzacott, J.A., "A Recursive Algorithm for Finding Reliability Measures Related to the Connection of Nodes in a Graph," Networks, Vol. 10, pp. 311-327 (1980)

Buzacott, J.A., "A Recursive Algorithm for Directed-Graph Reliability," Networks, Vol. 13, pp. 241-246, (1983)

Johnson, R., "Network Reliability and Acyclic Orientations," Networks, Vol. 1984, pp. 489-505, (1984)

Evans, T. and Derek Smith, "Optimally Reliable Graphs for Both Edge and Vertex Failures," Networks, Vol 16, pp. 199-204, (1986)

Ma, Y.W. and C.M. Chen, "The Application of the Random Graph Model for the Reliability Analysis of Dynamic Computer Networks," Proceedings INFOCOM, pp. 43-48, April (1984)

Provan, J. Scott, and Michael O. Ball, "Computing Network Reliability in Time Polynomial in the Number of Cuts," Operations Research, Vol. 32, pp. 516-526, (1984)

Rai, Suresh, "A Cutset Approach to Reliability Evaluation in Communication Networks," IEEE Transactions on Reliability, Vol. R-31, pp. 428-431 (1982)

Satyanarayana, A., "A Unified Formula for Analysis of Some Network Reliability Problems," IEEE Transactions on Reliability, Vol R-31, pp. 23-32, (1982)

Satyanarayana, A., and Mark K. Chang, "Network Reliability and the Factoring Theorem," Networks, Vol. 13, pp. 107-120, (1983)

Satyanarayana, A. and Jane N. Hagstrom, "Combinatorial Properties of Directed Graphs Useful in Computing Network Reliability," Networks, Vol. 11, pp. 357-366 (1981)

Satyanarayana, A. and A. Prabhakar, "New Topological Formula and Rapid Algorithm for Reliability Analysis of Complex Networks," IEEE Transactions on Reliability, Vol R-27, pp. 82-100 (1978)

Shier, D.R. and D.E. Whited, "Iterative Algorithms for Generating Minimal Cutsets in Directed Graphs," Networks, Vol. 16, pp. 133-147, (1986)

Wilkov, Robert S., "Analysis and Design of Reliable Computer Networks," IEEE Transactions on Communications, Vol. COM-20, pp. 660-678, (1972)

Willie, Randall R., "A Theorem Concerning Cyclic Directed Graphs with Applications to Network Reliability," Networks, Vol. 10, pp. 71-78 (1980)

SHORTEST PATH

Denardo, Eric V., and Bennett L. Fox, "Shortest-Route Methods: 1. Reaching, Pruning, and Buckets," Operations Research, Vol. 27, pp. 161-187 (1979)

Dial, R., F. Glover, D. Karney, and D. Klingman, "A Computational Analysis of Alternative Algorithms and Labelling Techniques for Finding Shortest Path Trees," Networks, Vol. 9, pp. 215-248, (1979)

Glover, Fred, Randy Glover and Darwin Klingman,
"Computational Study of an Improved Shortest Path
Algorithm," Networks, Vol. 14, pp. 25-36, (1984)

Pape, U., Algorithm 562: Shortest Path Lengths, ACM Trans.
Math. Software, Vol. 5, pp. 450-455, (1980)

Shier, D. R., "On Algorithms for Finding the k Shortest
Paths in a Network," Networks, Vol. 9, pp. 195-214 (1979)

Syslo, Maciej M., Narsingh Deo and Janusz S. Kowalik,
Discrete Optimization Algorithms, Prentice-Hall, Inc.,
Englewood Cliffs, New Jersey, 1983

ACKNOWLEDGEMENTS

I would like to acknowledge the support and assistance of the Air Force Systems Command, Air Force Office of Scientific Research, ESD/XR and ESD/MD. Without their support, this research would not have been possible.

In particular, there were a number of individuals who were extremely helpful to me and to whom I owe great thanks.

In ESD/MD, Mr. George Richardson shared his considerable expertise in the engineering problems found in the design and operation of communication networks. Lt. Col. Ted Mervosh made it possible for me to enrich my own professional background and he and Col. Richard Paul provided strong support.

In the MITRE Corporation, I would like to thank Mr. John Kattar, Dr. Rajan Varad, Mr. Burt Noyes and Dr. Bill Collins for lending me their vast experience and knowledge.

I also owe great thanks to Dr. Michael Ball, who provided 2000 lines of clear code for his algorithms, saving me a great deal of work.

RESEARCH INITIATION PROGRAM (RIP)

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Bolling AFB, Washington, D. C.

Conducted by the
Universal Energy Systems, Inc.

FINAL REPORT

Comparison of Testability Analysis Tools for USAF

Prepared by:	Dr. Sundaram Natarajan Bradley K. Herman
Academic Rank:	Associate Professor Graduate Student
Department and	Center for Electric Power
University:	Tennessee Technological University
Research Location:	Tennessee Technological University
Date:	February, 1991
Contract No:	F49620-88-C-0053/SB5881-0378

COMPARISON OF TESTABILITY ANALYSIS TOOLS FOR USAF

by
Dr. Sundaram Natarajan
and
Bradley K. Herman

ABSTRACT

As electronic systems become more complex, testability is becoming an increasingly important design consideration. To incorporate testability properly during the design phase, an appropriate testability analysis tool must be selected. In this report, we address the necessity for including testability in the design phase and develop a set of criteria to compare testability analysis tools. We analyze several tools with respect to the established criteria. Finally, we conclude with the summary and the recommendations regarding the selection of a suitable testability analysis tool for design for testability (DFT).

Acknowledgements

We thank the Air Force Office of Scientific Research for sponsoring this research. We also acknowledge Universal Energy Systems and the Center for Electric Power for their help and concern in all the administrative aspects of this research.

We also wish to thank the following individuals for contributing their time and products to our research: Mr. James Patterson of Automated Technology Systems corporation for providing I-CAT, Mr. Melvin Nunn of Naval Ocean Systems Center for providing CAFIT, Mr. Ralph DePaul of Detex Systems, Inc. for providing STAT, Dr. Randy Simpson of Arinc Research Corp. for the work associated with STAMP, Mr. Gene Biagi of BITE, Inc. for providing ASTEP, and Mr. Dean DeMeyer of WRALC for providing the circuits. Without their cooperation and contributions, this research would not have been possible.

I. INTRODUCTION

Recent USAF programs have sought to address the enhancement of both built-in and off-line testing of new systems early in the acquisition phase. This requires the inclusion of testability concepts in the design phase. It has the goal of eliminating interim contractor support (ICS) and thus reducing the life-cycle cost (LCC) of the system. In addition to reducing the LCC of the system, the incorporation of testability in the design phase also increases the reliability, maintainability, and availability (RM&A) of the system. The potential monetary savings and the increased quality of the system make the inclusion of testability in the design phase very attractive. Testability must be included at the lowest level of design to be effective. For electronic systems, this means that the testability criteria should be applied at the circuit card assembly (CCA) level. This is becoming increasingly more important with the increased CCA complexity of modern electronic systems. The task of including testability in the design phase of the CCAs lends itself to computer-aided design (CAD) tools. To ensure an accurate verification of testability requirements, a proper CAD tool must be chosen that is suitable for the job at hand. The previous studies of testability analysis tools have dealt mainly with system level analysis, but not with the board level analysis required by modern electronic systems. The other board level studies were performed before most of today's CAD tools were even in existence. Also the growing number and added maturity of CAD tools requires a new study to be performed.

In this report we first establish the objectives of our research and provide support for the research in the form of an overview of the design for testability philosophy. We then state and support the criteria and method by which we evaluate each testability tool. A section is devoted to each tool that briefly describes how each software meets the established criteria. These descriptions should not be considered complete overviews of the tools. For details beyond the scope of this report, any interested

individuals should contact the individual vendors. We conclude with a summary of the research and our recommendations.

II. OBJECTIVES OF THE RESEARCH EFFORT

The objective of our research was to analyze and compare several testability analysis tools as to how they perform a board level testability analysis of electronic systems. This involved evaluating several aspects of each tool from ease of input to clarity and accuracy of the outputs. Another main goal of this research was to determine one or more Figures of Merit(FOM) that can be used to set testability requirements. Since there is no standard method for obtaining an FOM, the resulting FOM from each tool is unique. The proper FOM would be one which describes the testability of a CCA in the fewest possible parameters with the most insight into its testability. To get an accurate evaluation, the CAD tools were tested with a variety of circuits, namely digital, analog, and hybrid circuits. For our study to be realistic, it was carried out using actual CCAs selected from the E-3 AWACS aircraft.

III. DESIGN FOR TESTABILITY

Testability addresses the issues involved in achieving the inherent ability in a system to be testable so that desired levels of fault detection and isolation can be achieved in an accurate, timely, and cost-effective manner. The testing of an item can be influenced by the physical design of the item, the electrical design of the item, and procedures and equipment used to test the item. With the introduction of sophisticated test equipment, lacking physical designs can be overcome with guided probe and bed-of-nails techniques. However, this becomes ineffective in test cost and time required to test. Even with the sophistication of the test equipment, the system must still be designed in such a manner that it is testable. This is known as being inherently testable. As systems become more costly and more complex with higher part

density on each CCA, the ability to test and maintain these systems becomes increasingly difficult. Only if a system is highly testable can confidence be placed in the fact that it is fault free. Once a system is known to contain a fault, the ability to isolate the fault is directly related to the inherent testability of the system. The testability of the system is one of the leading factors influencing the LCC of the system. This is because the maintenance of a system is influenced by its testability and it can account for 50% of its LCC.

The most common method of system design places the test engineer at the end of the design phase. Typically, at this point, the hardware has been designed and is at some point of production and the test engineer must find a way to improve the testability of the system through probe tests and other intrusive methods. This method is time consuming and quite often the test requirements cannot be met due to a lack of inherent testability of the design. The decision must then be made whether to redesign the hardware or relax the test requirements. Since the hardware is being produced in some fashion at this point, the redesign of the hardware would be very costly and time consuming. On the other hand, the relaxation of the test requirements would reduce the testability of the system and thus increase its maintenance cost and hence its LCC. The most logical solution to this is to include testability in the initial design phase so that test requirements can be met before the system enters the production phase. The test requirements can then be treated as additional design requirements for the system. This is called design for testability (DFT). Addressing testability issues in the design phase enhances the testability of a system thus reducing its LCC.

As the complexity of the system increases, the ability of a designer to incorporate testability by the old time-consuming manual methods becomes impossible. Thus MIL-STD-2165 becomes more of a guideline than a method of analyzing testability. At the early stages of the design phase, the designs can change on a daily basis. Therefore, some method of incorporating these changes rapidly into the testability analysis becomes a necessity. This

problem lends itself quite well to computer-aided design(CAD) systems. Therefore CAD tools have become necessary for proper testability analysis. There are several such tools available. Some of the popular ones are: CAFIT, I-CAT, STAMP, ASTEP, and STAT. Our main concern was to compare these tools for the USAF. ASTEP was considered in the original proposal, but was removed from this study by request because it is primarily a system level testability analysis tool and it is not usable for testability analysis at the CCA level.

IV. CRITERIA FOR TESTABILITY ANALYSIS TOOLS

The first criteria to evaluate was the input requirements of each testability tool as to how much effort is required to prepare the design information in a form suitable for the software. This involved evaluating the modeling method used for each program. Other aspects such as the flexibility of the modeling procedure and the time required to do the modeling should also be included in the criteria. The next criteria evaluated was the ease with which the modeling information can be entered into the program. This included the user interface as well as the form in which the information is presented back to the user. The data should be presented in such a way as to allow the user to debug the information quickly. The error checking ability and the clarity of the error messages presented to the user are two criteria that must be addressed for tasks of the size being considered (i.e. large CCAs such as the ones we used). The run-time of the programs is also very important since one would not wish to wait overnight for the analysis of a single CCA.

The primary criteria for the evaluation of the testability analysis tools is the information in the form of a FOM from the output of each tool. The output should be concise for the quick assessment of the inherent testability of the CCA. It should also be detailed enough to provide insight into the details of the testability in a specific portion of the CCA. This information should include such aspects as the various ambiguity groups,

feedback loops, types of testing strategy to apply, etc. In addition to the quantitative outputs, some qualitative outputs such as suggestions to improve the testability are also desirable. These include where to break feedback loops, where to add test points, test points to eliminate, and any other suggestions.

In addition to these criteria, some other aspects of the individual programs were also evaluated. These were primarily features regarding the testability analysis unique to a particular program.

Finally, to summarize the evaluation criteria, they are in the following order of importance:

- (1) A concise FOM yielding a measure of testability
- (2) Suggestions to improve testability through the breaking feedback loops
- (3) Suggestions to improve testability through the addition or elimination of test nodes
- (4) Information on ambiguity groups and feedback loops
- (5) Testing strategy or procedure
- (6) Data input requirements
- (7) Amount of modeling required
- (8) User interface.

V. METHOD OF EVALUATION

Our primary concern in the evaluation of the testability tools was how the tools meet the criteria established in the previous section. The three circuits chosen were selected to give a feel for how the tools met the criteria on a variety of real-world electronic circuits. Another criterion that was used in selecting the circuits was the experience of the WRALC personnel in developing the depot for these circuits. The actual numerical results are of little consequence since each circuit, that any user will analyze, will have its own unique characteristics. How the tools met the challenge of evaluating different types of circuits was of a greater importance. Since every circuit will yield unique results, the numerical results themselves cannot be used to

evaluate the tools. Each tool was scored on a scale of 0 to 10 as to how effectively each of the criteria was met. Each criteria was assigned a weighting factor based on its importance to design for testability (DFT).

The first criteria is the most important since it characterizes and summarizes the testability of the CCA and it received a weight of 10. The second and third criteria were both given weights of 9 since they are the primary means of improving testability through design changes. The fourth criteria was assigned a weight of 8. This is because the detailed information can be used to improve testability even though no direct suggestions are made. The fifth criteria is more limited in the information that it can provide with regard to the design procedure so it received a weight of 6. The last three criteria do not deal directly with the testability analysis and the design of a CCA. However, they are still important because even an effective design tool may not be practical in its use. Thus, each of the last three criteria was given a weight of 5.

Once all the scores were assigned, each tool had its weighted total score calculated from the assigned scores and weights. Any evaluation is bound to be subjective, but we believe that our independence from any government agency or private corporation limits the subjectiveness as much as possible.

VI. CAFIT

CAFIT stands for Computer-Aided Fault Isolation and Testability Model. CAFIT was developed by ATAC of Mountain View, CA. It is a government owned testability analysis tool. CAFIT was designed to enhance the testability of electronic circuits, both digital and analog. It provides outputs, after performing several forms of analysis, that aid in designing circuits to be inherently testable. For this study, CAFIT was run on an IBM-PC AT compatible. It requires an EGA compatible monitor and occupies approximately 2 Mbytes of hard disk space. An additional 1.4 Mbytes of disk space is taken up by the library, but this will vary depending on the

number of device models used. It can also be run on Mentor-Graphics workstations.

CAFIT provides a library utility that builds and maintains a model library of parts. This library contains all of the logic, attribute, failure rate, and pin information for each device in the library. The user can modify and add devices easily to the library with the help of the reference manual. Digital devices can be entered easily from the data sheet of most devices with slight modifications into the format required by CAFIT. For the digital devices, the model contains the actual function of the device in terms of HIGH, LOW, TRI-STATE, etc. For analog devices, no functional information is used. Thus the analog model is purely topological. This mixture also allows for the modeling of hybrid components such as D/A and A/D converters. Care must be taken when using analog devices to ensure that the proper dependency model is used.

CAFIT is designed to accept schematic information from a netlist provided by a CAE package such as the CT-2000 software package from Case Technologies, Inc. If such a package is not available, the connectivity information from the schematic can be entered into CAFIT through an ASCII file in the proper format. We used this method since the CAE tool was not available for our use. CAFIT takes this information along with the dependency model information previously entered to do the testability analysis. If the CAE system is available, results are shown in color format on the schematic. Otherwise the output files provide the means to view the results. Input for the ASCII file is easily obtained directly from the schematic and this does not require any special knowledge of the inner workings of the circuit if the circuit is purely digital. However, if analog components are used, then the knowledge of the operation of the circuit is essential to get the proper topological models. The only constraint on the labeling of nodes is in the labeling of test nodes, input signals, and output signals. Test node labels must have a TS prefix while signal input and output signals must have a SG prefix. Overall, CAFIT requires very little preparation of the input information if the models have

already been created within CAFIT. This will most often be the case once a library of devices has been established.

Performing a basic testability analysis with CAFIT for an existing design is relatively easy. A basic analysis attempts to locate structures inherent in the design that will inhibit the testing of the design. These structures include negative reconvergence, logic redundancy and, most importantly, the information on feedback loops present in the design. Negative reconvergence is a design condition in digital systems where a portion of logic prevents an input from controlling an output. This condition is not necessarily a design flaw, but it can seriously impede the testing of the circuit. Logic redundancies, which inhibit the testing of the circuit, are also identified. With logic redundancy, the observability of the circuit is reduced. The feedback loop portion of the report is probably the most useful. This provides a detailed output listing of all the components involved in each loop as well as nodes at which to break each loop. The next type of analysis provided by CAFIT allows the user to select parameters about the test nodes and perform a detailed testability analysis with regard to signal controllability and observability.

If the test nodes have already been selected, an analysis can be done to determine signal coverage with only these nodes. The signal coverage of a device is a measure of whether or not stuck-at faults in device can be tested for. This is a function of controllability, observability, and the number of tests permitted. Controllability and observability figures using the signal input and output and test sites are given along with the signal coverage for each device. In addition to the selected test nodes, an additional set of nodes can be chosen by the user. CAFIT then selects the nodes that will yield the highest signal coverage. If 100% coverage can be achieved with fewer test nodes, then the lowest number of nodes are chosen. The test node selection is probably one of the most powerful aspects of the analysis that CAFIT provides. This allows the user to experiment with various configurations to determine the optimal number of test nodes for a

given CCA. The number of test signals is an additional parameter that can also be varied. This constraint is becoming less prevalent as sophisticated test equipment becomes commonplace. This is also more applicable to digital circuits where CAFIT has the knowledge of the various states of the circuit from the model description. Analog circuits may require only one or possibly several tests to be performed on a single node, depending on the circuit.

The outputs from CAFIT can be used for the selection of test nodes, the selection of points to break feedback loops, and the estimation of controllability and observability. However, most testability requirements are in terms of fault isolation parameters and ambiguity group sizes. Even though a node is controllable and observable, it may not have a unique test that identifies only that node, leading to the creation of ambiguity groups. This is especially true when dealing with feedback loops. The entire loop is an ambiguity group even though each device may be controllable and observable. This is especially important in the testing of analog and hybrid circuits where feedback networks are invariably used. The information provided through controllability and observability is not enough to describe the testability of an analog or a hybrid circuit thoroughly.

The user interface for CAFIT is one of the easiest to use and understand. It is designed in such a way that even an inexperienced user can use the program. On screen help may be called up throughout the program. CAFIT provides information on the screen without cluttering it up. The error messages provided were descriptive and aided in the debugging of the various circuits tested. The only shortfall in the user interface is having to prepare an ASCII file if the CAE system is not available. It is acknowledged that CAFIT was designed to be used with an accompanying CAE system, but a provision for entering the schematic information directly into CAFIT and having CAFIT manage the files would have been helpful. CAFIT also writes the output data to one file only without the option to change the name of the file. Thus, after each run, it overwrites the existing file making it necessary to continually rename the output file after each run. This is not

a serious drawback, but it can be frustrating after waiting 45 minutes for results and then realizing previous results were accidentally overwritten.

CAFIT is primarily geared toward the analysis of digital circuits. It can analyze analog and hybrid circuits also. However, the terminology used by the program and the user's manual suggest that the analog portion has been added as an afterthought. The lack of fault isolation and ambiguity group parameters limits the usefulness of the output to verify whether the desired levels of testability are being met. However, the test node selection and feedback loop breakpoint analysis make CAFIT a tool for choosing test nodes and making changes in the circuit to increase its inherent testability.

CRITERIA and SCORES for CAFIT:

- (1) 5 : The controllability and observability figures alone are not enough to fully characterize the testability of a circuit.
- (2) 8 : The suggestions for selecting feedback loop breakpoints are essential.
- (3) 8 : The time spent selecting appropriate test nodes is greatly reduced with CAFIT suggesting the most effective test nodes.
- (4) 3 : The information on feedback loops is helpful, but there is a total lack of information on ambiguity groups.
- (5) 0 : No outputs concerning testing procedure.
- (6) 5 : The existence of the interface with the CAE tool can make data input simple. However, the alternative of having to format an ASCII file with all of the circuit information is tedious.
- (7) 6 : Minimal modeling is required once a library of devices is established. However, modeling of analog circuits must still be done on an individual basis.
- (8) 8 : The user interface is simple and easy to use and has plenty of on-line help.

VII. STAT

STAT is a product of Detex Systems, Inc. of Orange, CA. STAT, along with its predecessor LOGMOD, is a testability analysis tool designed to enhance the testability of systems. This applies to the enhancement of the testability of new systems through their design and for the maintenance of existing systems. Our study will focus on the design aspects of its capability. STAT runs on a variety of PC platforms, but a minimum of a 286-based PC is recommended with a sufficiently large hard drive to store the STAT program and all of the user's databases. It requires approximately 2.5 Mbytes of disk space for the program itself. The basis for the analysis that STAT performs is the dependency model information supplied by the user.

A dependency model is one that simply shows the relationships (dependencies) between the various nodes and items in a system. This can be interpreted as describing the information flow within a general system and as a signal flow for an electronic system. The topological information from a circuit is entered into STAT completely in the form of dependency models, not as the connections between the actual electronic components. This type of modeling has advantages as well as disadvantages. The main advantage for electronic circuits is that it bridges the gap between the descriptions of digital and analog circuits. Another advantage of the dependency modeling for electronic circuits is that it provides considerable flexibility to describe a particular circuit. This includes what aspects of a particular signal are important as well as the level of modeling for each portion of the circuit. For example, it allows the user to choose whether or not to consider packages containing more than one device to be considered as individuals or as a single item. Being able to describe the particular aspects of one signal as individual signals is particularly in analog circuits where biasing and actual signal information must be considered separately. The disadvantage of this type of modeling is that it requires an intimate knowledge about the operation of a particular circuit in order for it to be

accurately modeled. Modeling a complex analog or hybrid circuit can be a time consuming and frustrating task even when the circuit is familiar to the user. However, the proper modeling up front will help to reduce future problems and the time required to perform a proper testability analysis using STAT.

The entering of the dependency model information into STAT is very straight-forward. The information is entered in an interactive fashion with STAT updating and maintaining the databases transparently. Each node in the dependency model is entered as a test node. One aspect of STAT, that is particularly frustrating, is that all node labels must have a 'T' prefix and all item labels must have an 'I' prefix. This requires the user to maintain some type of translation table independently. Being able to use descriptive labels would aid in interpreting the results of the testability analysis. Descriptions can be added to the dependency information as it is entered. The dependency model information is displayed back to the user in a manner that allows errors to be seen visibly and corrected. Once the basic dependency information has been entered, the model can then be processed to find and correct any additional errors. The error messages supplied are descriptive and they can be corrected easily.

The evaluation of a model with STAT is not a one step process, but rather a series of processes. When a model is first processed, all nodes are considered as test nodes. This becomes the base model for all subsequent evaluations. This is the point at which STAT allows the user to perform a variety of "What if?" type scenarios. This is done by first creating a new model that is identical to the base model with a simple keystroke. The user can then modify the characteristics of each node within the model. Most nodes will be chosen as non-test nodes. However, the aspects of the chosen test nodes can also be further modified. The type of access required to reach a node, such as external or probe accessible, can be specified. The cost and time to test each point can also be specified. The failure rate and replacement time and cost of individual items can also be modified. Other factors can also be modified and weighted. The weighting allows the user to perform an

analysis based on individual specifications.

One aspect of the operation of STAT deserves special mention. Along with identifying feedback loops within the model, STAT has an interactive feature called the Feedback Loop Breaker. This feature identifies feedback loops, then recommends the nodes at which to break the feedback loops and the new characteristics resulting from the breakage. The resulting ambiguity groups can be viewed immediately. This allows the user to continue to modify the loops until the required ambiguity group requirements are met. This is just one aspect of a very workable user interface. Once all the terminology associated with STAT's modeling is understood, it is easy to maneuver through many screens and perform a testability analysis.

The outputs from STAT take on a variety of forms that range from the concise and informative to the long and drawn out. The longer outputs are more applicable to the maintenance aspects of testability and not to the design process. Each output report contains a summary section and a detailed section. This is very helpful when trying to locate a particular piece of information without turning through perhaps hundreds of pages. The management model report contains a graphical representation of the topological dependencies for a particular model. This allows the user to more clearly see the loops that are not easily identifiable in the original circuit diagram. The feedback loop indicator report contains a detailed information on all feedback loops in a model. This includes information on the relative complexity of the feedback loops and information on the sizes of the ambiguity groups resulting from the breaking of the loops. The topological indicator report contains the most information to determine the inherent testability of a circuit. The fault isolation level is given for each ambiguity group size present in the circuit. This FOM is a primary means of verifying whether testability requirements are being met. All the characteristics of each ambiguity group, test, and item are given in the detail section. This information is in an easy to read form, but for large circuits, the detailed section can become quite large. The last report produced is the fault

isolation indicator report. The processing of this report provides a means for determining the number and types of tests to use. The user has options regarding the acceptable size, time to replace, and cost to replace for ambiguity groups. The user can also select the number and type of tests to consider. The report then lists the suggested tests to use and tests that are not used. This report also includes a detailed flow diagram for the isolation of faults. This diagram includes the cost and time spent testing to reach a particular point in the isolation procedure. The first three reports, namely the management model, feedback loop indicator, and topological reports, are better suited to evaluate the testability of a particular design.

Using the dependency modeling, STAT can analyze any circuit regardless of whether it is analog, digital, or hybrid. The only difference is in the amount of modeling required for a particular CCA. The multitude of output information allows the evaluation of just about any aspect of the testability. The most important and useful of which are the detailed fault isolation and ambiguity group parameters.

CRITERIA and SCORES for STAT:

- (1) 9 : The presentation of the fault isolation levels in a simple table is the ideal result to express the testability of a circuit.
- (2) 10 : The feedback loop breaker provides information directly on the results of breaking loops. The interactive nature of this feature makes dealing with feedback loops extremely easy.
- (3) 9 : Provides informative, detailed suggestions regarding the placement of test nodes.
- (4) 9 : STAT provides very detailed information on all aspects of ambiguity groups and feedback loops.
- (5) 9 : Very detailed descriptions of test strategy and suggestions as to what type of test to use.
- (6) 6 : The interactive data editor makes data entry easy, but

having to have the proper prefixes for all nodes and items is irritating.

- (7) 4 : The development of the dependency model for complex circuits can become very tedious and time consuming.
- (8) 8 : The descriptive names of all the choices presented and the on-line help make STAT an easy program to operate.

VIII. STAMP

STAMP, which stands for System Testability and Maintenance Program, is a tool designed for the analysis of system testability and fault-isolation strategies. STAMP is a program from ARINC Research Corporation of Annapolis, MD. It runs on an HP-1000 mini-computer and recently a PC-based version has been announced. STAMP is for the use of ARINC employees only and primarily for government contracts. It is unfortunate that, at this time, it is not available to the end user directly. For a tool to be effective in the design phase, it must be readily available to the design personnel. Then, when design changes in the circuit occur, their effect on the circuit's inherent testability can be quickly determined and suggestions made back to the design personnel on how to improve the circuit's testability. We developed the models for the three circuits and sent them to ARINC to be processed. The quick turn-around for receiving the results would be sufficient when doing an analysis of a previously designed circuit for testing purposes only. However, for a design work, the results must be available as quickly as possible. STAMP performs its analysis based on the first-order dependency model.

The discussion on the dependency model in the previous section applies to STAMP as well except for the differences in terminology for tests and items. STAMP refers to these as events and elements, respectively. The use of the dependency model is essential for modeling hybrid and analog circuits properly. Again the main disadvantage of dependency modeling is the large amount of time required to model analog and hybrid circuits. The user has several options for inputting the dependency information. The dependency

information can be entered directly into STAMP through an interactive interface, through an ASCII file, or through the wire-list output of the popular schematic drawing program Schema. We produced our dependency lists which were then entered directly into STAMP manually. STAMP has default labels for all the aspects of the dependency model. However, these labels may be modified so that the labels on the output information can be related directly back to the original schematic. In addition to the first-order dependency information, a wide range of modifiable parameters are also available. These include failure rates, different types of test, test time and cost, test grouping, component grouping, etc. The ease with which the data can be entered is not a criteria that we can judge fairly since we did not actually run STAMP. Likewise, the options available when obtaining output cannot be fairly judged. However, the extensive outputs can be analyzed and they do contain the information on testability parameters.

The outputs are separated into two groups, namely, the testability analysis and the fault-isolation analysis. In addition, the lists of all components, dependencies, and a wide variety of other lists are also available from the output. The amount of paper consumed can become excessive, but the reports can be printed separately. Our primary interest is in the testability analysis section which provides 24 different measures of testability. All the results are normalized to provide the user with an idea of how the measures compare relatively. Most of the measures are slightly esoteric, but several provide very information on what to concentrate and on how to improve the inherent testability. The outputs, that are the most informative, are the operational isolation measures. These are commonly referred to as fault-isolation levels. These are one of the means of stating the required testability specifications in the SOW of government contracts. They are the primary FOM for inherent testability. The two primary results give the fault-isolation for a given ambiguity group size. One output considers the weighting of the failure rates while the other does not. This allows for additional flexibility for the specifications to achieve desired testability levels. STAMP

provides information on ambiguity groups such as number, size and component members. One type of output, that is worth noting, is the listing of excess tests and redundant tests. This information is very useful when the number of test nodes available is limited by some other factor such as connector pins. Components and tests that comprise feedback loops are also listed. The user's manual states that, in addition to listing the loops, additional tests are identified that could be in dealing with the loops. We could not find these suggestions in any of the outputs that were supplied to us. These are the most useful of the testability analysis outputs. Additional outputs are available that describe a variety of other conditions and parameters.

The fault-isolation analysis provides the information that a technician would use in the isolation of a fault. Lists are given which identify the tests that will be bad given a failure in a particular component. Probabilities of failure figures are also given for all components and tests. A tabular decision tree and a graphical tree can be produced that detail the steps a technician would follow in the fault isolation procedure. The procedure can be affected by different weightings and options associated with test time, test cost, test groupings, component groupings, etc. Finally, a summary of the testing procedure in terms of the number of steps required to isolate faults is given.

With the use of dependency modeling, STAMP can also analyze any type of circuit regardless of whether it is digital, analog, or hybrid. The drawback to this is the extensive amount of modeling that must be performed to properly analyze analog and hybrid circuits. The major feature missing from STAMP is the lack of information on breaking up feedback loops. The most useful output provided is the data on fault isolation levels related to ambiguity group size. The extensive output information for about every conceivable seems excessive, but may find application in different situations.

NOTE: STAMP will only be an effective tool for incorporating testability in the design phase if it is available for

use by the design personnel. The time needed to obtain the results and the need to experiment with many different testing scenarios makes this a necessity.

CRITERIA and SCORES for STAMP:

- (1) 10 : The operational isolation measures for both the consideration or exclusion of failure rates is the most result.
- (2) 0 : No suggestions are provided for the breaking of feedback loops. This is a major drawback since the existence of feedback loop is a primary contributor to poor testability. Invariably all real world CCAs will have feedback loops.
- (3) 6 : The listing of redundant and excessive tests is very useful for large designs where all possible paths to the output tests are not always clear. However, no suggestions are made regarding the placement of additional tests.
- (4) 9 : STAMP provides detailed listings of every aspect of the dependencies, ambiguity groupings, etc.
- (5) 8 : Detailed test procedures are given in decision tree form.
- (6) 8 : The existence of a CAD interface with Schema can make data input easier. The flexibility of the labeling of components and tests makes data entry and verification easier.
- (7) 4 : The development of the dependency model for complex circuits can become very tedious and time consuming.
- (8) N/A : We cannot evaluate this aspect of STAMP since we did not actually use the program.

IX. I-CAT

I-CAT, which stands for Intelligent Computer Aided Testing, is a model based expert system designed for fault-diagnosis applications. I-CAT is a software program from Automated Technology

Systems Corporation. We ran I-CAT under SunView on a Sun SPARCstation 1 with 8 MBytes RAM installed and a color monitor. Although I-CAT is primarily designed for fault-diagnosis, it has several features that make it attractive for performing testability analysis. I-CAT also requires essentially the same type of models as those required by STAT and STAMP. The only difference is that I-CAT refers to the model as a functional model rather than a dependency model, although both contain the same information.

The model information is entered into I-CAT through a graphical interface. I-CAT defaults to starting all component labels with a 'C'. This can be altered to anything that is considered appropriate by the user. When test nodes are placed in the diagram, they can have any label that has not already been used. Since the model can be seen, it is easier to debug the input data with I-CAT than other programs. However, this is not always the case with large, cumbersome models. Even when using the grid to aid in the alignment of connecting lines, it is often difficult to make the connections if the view is not zoomed in. Any open lines resulting from these bad connections have the same effect as in a real circuit since the signal path is broken. These are sometimes difficult to spot unless the view is zoomed in quite close. Trying to maneuver around a large model can be frustrating due to the lack of a dynamic scroll. However, I-CAT encourages the use of nested models to make the model easier to create, edit, and view. Any component can contain a substructure that can consist of another model. I-CAT contains a built-in syntax checker that verifies the graphical model for connectivity errors. This is essential since the graphical interface can sometimes be deceiving. Along with the functional model, various other parameters may be entered, such as component cost, component failure rate, and test cost.

I-CAT's development tools generate a wide variety of outputs that have many uses. We focus primarily on the testability analysis data. Several pertinent input options are available to the user that have an effect on the testability analysis, namely, whether or not to allow multiple faults, whether or not to use only the assigned test nodes, and whether or not to test inside all the

substructures. If the assigned test nodes are not used, the number of test nodes to use may be entered. The primary outputs of the testability analysis are the line tree and analysis report. The line tree provides a graphical representation of the testing procedure. This is also given in other forms such as a bar tree and a flow chart. The analysis report contains statistical information on the cost of diagnosis, the replacement costs, and the ambiguity group sizes. Statistics such as variance and standard deviation are given. The ambiguity group information contains no information about which components belong to which group. This makes it difficult to modify a design to have smaller ambiguity groups. Information is also provided on feedback loops and test nodes. I-CAT identifies feedback loops and the components that comprise each loop. No suggestions are given about where to break the loops. The test node data is the most useful of the testability analysis outputs. It details which nodes are used, not used, and any additional nodes that may be required. The last output is only given if the assigned test nodes are ignored. It is not clear what testability parameter of the model is being maximized by the choice of test nodes. There are no outputs that give the testability of the model in terms of fault coverage and ambiguity group sizes. These two aspects are the important ones for DFT and it is unfortunate that these are not available from I-CAT. I-CAT can also produce diagnostic programs in various languages, test requirement documents (TRD), netlist information, and a Personal ATLAS Workstation (PAWS) data base.

I-CAT also has several additional features that deserve mention, although not directly related to design for testability. These are a troubleshooter, fault simulator, and a rule editor. The troubleshooter is an interactive graphical interface for diagnostics. The user enters components known to be good and the results of tests. The troubleshooter can direct a technician on how to proceed with the testing of a device. This is all accomplished through extensive dialog boxes that are full of information previously entered into I-CAT. The symptoms reported at test nodes can also be entered into I-CAT. The user can even step backward

through a diagnostic session. Perhaps the most important aspect of the troubleshooter is that the failure information can be saved to create a history of components that fail. The fault simulator can be used to compare the testing strategies used by technicians and the strategies offered by I-CAT. I-CAT randomly introduces a fault into the model and it is up to the user to find its location. The user can keep track of the number of steps required to isolate a fault and then let I-CAT find the fault. This way a technician can train on a particular circuit without having to damage the actual hardware. The rule editor is probably what distinguishes I-CAT from the other testability tools. Information on all aspects of the tests and the testing procedure is entered through the rule editor. The parameters associated with tests range from the dimensions of the measured quantities to the probabilities of test results. These quantities along with the failure history of the model from the troubleshooter can be used to increase the quality of the testing procedure. These features are primarily concerned with manual testing or any type of testing where operator intervention is necessary.

With functional modeling, I-CAT can also analyze all types of circuits whether they are analog, hybrid, or digital. Once again the drawback to this is the extensive amount of modeling required for analog and hybrid circuits. Although I-CAT is primarily a fault diagnosis tool, the information on test node selection and feedback loops is informative. However, the lack of fault isolation parameters and details about ambiguity groups make I-CAT better suited for establishing testing procedures once the circuits are designed.

CRITERIA and SCORES for I-CAT:

- (1) 0 : No final testability measure is given.
- (2) 0 : No information on where to break feedback loops.
- (3) 9 : The details concerning test nodes are the most useful of all the outputs with regard to design for testability.
- (4) 4 : Only limited information on ambiguity groups and feedback

loops.

- (5) 8 : Extensive outputs concerning the testing of the circuit including source code and documents.
- (6) 8 : Despite the sometimes picky graphical interface, I-CAT is flexible with labels and the model data is easy to enter and verify.
- (7) 4 : The development of the functional model for complex circuits can be tedious and time consuming.
- (8) 9 : The menu driven, mouse controlled interface was easy to use and very self-explanatory.

X. SUMMARY

There are sufficient motivations to include testability during the design phase of electronic systems. To include testability analysis properly, the appropriate tool must be chosen to evaluate and improve the testability of the CCAs. The testability analysis tool must be able to evaluate circuits whether they are digital, analog, or hybrid.

Each tool was evaluated for a digital, analog, and a hybrid circuit to assess the function over a wide variety of possible applications. Each criteria was scored on a scale of 0 to 10 for each tool. The scores assigned to each tool for all the criteria are summarized below in Table 1 along with the final weighted total.

TABLE 1: Scores for testability tools.

TOOL	CRITERIA (WEIGHT)								WEIGHTHED TOTAL
	1(10)	2(9)	3(9)	4(8)	5(6)	6(5)	7(5)	8(5)	
CAFIT	5	8	8	3	0	5	6	8	313
STAT	9	10	9	9	9	6	4	8	477
STAMP	10	0	6	9	8	8	4	n/a	334*
I-CAT	0	0	9	4	8	8	4	9	266

* - STAMP was not available for our use. As a result the user interface could not be evaluated and the total score reflects this loss.

I-CAT is primarily a fault diagnosis tool and this is evident in its strong showing in the criteria associated more with this aspect of testing and not DFT. Its multitude of outputs concerning the testing of a CCA makes it ideally suited to repair depot development. I-CAT has the ability to learn from the experience of personnel with its use of expert systems technology. Its ability to lead a technician through the repair of a device also makes it attractive for use at the depot level. However, in its present form, it is not suitable tool for testability analysis during the design phase.

CAFIT has the advantage of being government owned and thus it is readily available for use. Its library facility reduces the time required in modeling of digital circuits by maintaining the model for each device. However, the complexities associated with hybrid and analog circuits make their modeling difficult and necessary on an individual basis. CAFIT can aid in the selection of test nodes, but does not give any information regarding the ambiguity groups, testing procedures, or an FOM dealing with fault isolation levels and ambiguity groups. If improvements can be made to CAFIT to make it more useful for the design of CCAs, it would be an attractive testability analysis tool since it is already government owned.

STAMP provides an output for just about every aspect of a CCA. STAMP does provide a concise FOM in terms of fault isolation levels and ambiguity groups. However, even if the total assigned point value was added for the user interface, STAMP still would fall short due to its lack of suggestions on the placement of test nodes and breaking of feedback loops. These features are required if the tool is to be effective in the design process. Furthermore, its lack of availability prohibits it from being an effective tool for design purposes.

STAT provides very extensive outputs that can aid in the design of testable CCAs. Its numerous user options allow the

designer to use a type of "What if.." scenario. The numerous test node suggestions and the interactive feedback loop breaker make STAT a suitable tool to aid in the design of testable CCAs. STAT also provides an FOM in terms of fault isolation and ambiguity groups. Its only drawback is the time consuming effort that is required to develop the dependency models. But if a facility to model electronic components could be established, the time required to develop the models can be eliminated.

XI. RECOMMENDATIONS

Based on the previous discussions and observations, we conclude that, at the present time, STAT would be the most appropriate testability tool to be used in the design phase. It possesses all the essentials to aid in the design of testable CCAs. However, if some improvements can be made to CAFIT, it can also become a powerful tool for DFT. If the latter is preferred because it is already government owned, the following suggestions are made for the improvement of CAFIT:

- (a) Provide an FOM in terms of fault isolation levels and ambiguity group sizes.
- (b) Provide information on ambiguity groups such as size and number.
- (c) Develop a model library for analog components.
- (d) Allow data to be entered easier if the CAE interface is not available.
- (e) Provide information on testing strategy or procedure.

At the present time, SMARTCAT, the upgrade to CAFIT, was not available for our use. It is not known whether any of these suggestions have already been incorporated. However, if the above improvements are made, SMARTCAT should be looked at closely to determine whether it is an acceptable tool for DFT.

REFERENCES

Natarajan, S. and B.K. Herman, "Analysis of Testability Concepts and its Application to RSIP," Under AFOSR contract F49620-88-C-0053, Aug. 18, 1989.

Bussert, J., "Testability Analysis Tools on a Military System," Technical Report TM-3143-1717, Naval Ocean Systems Center, Sept. 30, 1987.

Fritzemeier, R.R., H.T. Nagle and C.F. Hawkins, "Fundamentals of Testability - A Tutorial," IEEE Transactions on Industrial Electronics, pp. 117-128, May 1989.

Robach, C. and S. Guibert, "Testability Measures: A Review," Computer Systems Science and Engineering, pp. 117-126, July 1988.

Goldstein, L.H., "Controllability/Observability Analysis of Digital Circuits," IEEE Transactions on Circuits and Systems, Vol. CAS-26, No. 9, pp. 685-693, Sept. 1979.

Nagle, H.T., S.C. Roy, C.F. Hawkins, M.G. McNamer and R.R. Fritzmeier, "Design for Testability and Built-In Self Test: A Review," IEEE Transactions on Industrial Electronics, pp. 129-140, May 1989.

CAFIT User's Manual, ATAC, Mountain View, CA.

STAT, Detex Systems, Inc., Orange, CA.

STAMP User's Manual, ARINC Research Corp., Annapolis, MD.

I-CAT User's Manual, Automated Technology Systems Corporation, Hauppauge, NY.

"Protractive Integration of Testability in System Acquisition," RAC Newsletter, Jan. 1990.

FINAL REPORT TO UNIVERSAL ENERGY SYSTEMS, INC., DAYTON, OHIO

1989-1990 AFOSR RESEARCH INITIATION GRANT

ANOMALOUS EFFECTS OF WATER IN FIRE FIGHTING:
FACILITATION OF JP FIRES BY AZEOTROPIC DISTILLATION EFFECTS

William W. Bannister

Department of Chemistry, University of Lowell, Lowell, MA 01854

ABSTRACT

Water, when mixed with or applied to hydrocarbon fuels, substantially increases the rate of vaporization, with significant decreases in boiling points of the fuel components, due to azeotropic "steam distillation" effects. At ambient room temperatures it was found in this investigation that there is essentially no increase in volatility (upon addition of water to the fuel), in terms of reductions of flash points of fuel components, and therefore there is essentially no increased flammability for mixtures of fuel with water at ambient temperatures. It was also found, however, that the increases in volatility for water-fuel mixtures can be very pronounced in high temperature situations, as would be observed in fully developed hydrocarbon fuel fires which are being extinguished with water fog, AFFF, or other water based extinguishing agents or systems. Moreover, it was found in this investigation that the increased volatility effects are particularly severe for low volatility (high boiling point) fuels such as JP-8, and that can be expected to pose correspondingly severe problems in fire fighting efforts for such low volatility fuels.

INTRODUCTION

Water has been used for thousands of years as a fire extinguishing agent, either alone or as the main component of a variety of agent compositions such as AFFF or other water based extinguishing compositions. The greatest single effect of the water in such applications is the great cooling capacity of the water, which thus provides a capability for lowering the temperature of the burning fuel below its flash point, thus removing one of the four essential conditions for maintenance of the fire. (In addition to heat, the other essential bases of the so-called fire tetrahedron are oxygen, the fuel itself, and the existance of propagating free radical pathways in the flame system.)¹

It has been very well established that there are several types of situations in which application of water actually serves to intensify a fire.

1. When applied to very hot oil or grease fires, water will flash into steam, causing a spattering effect which will greatly intensify the fire.
2. Water reacts explosively with active metal fuels such as sodium.
3. Direction of a strong, vigorous jet of water from a fire hose into burning liquid fuel can result in mechanical "digging" effects which will scatter the burning liquid over a much wider area, causing a serious increase in size of the blaze.

4. "Boil over" can result from formation of a heat wave progressing downward through burning fuel floating on water, finally reaching the water and causing this to come to a rapid boil with forcible ejection of the burning fuel upward from the surface. "Boil over" is observed only for burning fuel mixtures comprised of both high and low density components (i.e., the effect is not observed for pure liquids); the fuel mixture must be floating on water; and the effect requires several hours of build-up time before it is observed.²

None of the above are in any way anomalous; none are involved in any way with the type of azeotroping effect to be described in this report.

It is curious that azeotroping have thus far never been considered in any firefighting research efforts, particularly since it can be shown to be at least as serious for fire fighting considerations as any of these other four effects. Azeotropy is certainly a well-known and well described phenomenon, forming the basis of what is variously termed "steam distillation", or immiscible phase azeotropy. This is a technique which has been practiced on a very large industrial scale, as a means of performing distillations at relatively low temperatures for what would ordinarily be very low volatility, high boiling point liquids. A brief discription of the principles of steam distillation is provided below. (For a more complete discussion of this effect, see references [3] and [4].)

As shown in Figures 1 - 3 for benzene, xylene and water, the boiling point of any liquid or mixture of liquids is that temperature at which the vapor pressure of the liquid system exactly equals the atmospheric pressure (the standard atmospheric pressure being 760 mm Hg). (Benzene and xylene will be discussed in detail in this paper, since each has boiling points which are analogous to the boiling points of volatile JP-4 and less volatile JP-8 fuels, respectively.)

In Figure 1, the boiling point of benzene is seen as 69° (176° F); at which temperature its vapor pressure is 760 mm (one atmosphere). In Figure 2, xylene boils at 139° C (282° F), with a vapor pressure of 760 mm; and water has a vapor pressure of one atmosphere at 100° C (212° F).

If two liquids are immiscible (insoluble in each other), each phase will exert its own vapor pressure at a given temperature, and the total pressure will then be the sum of the vapor pressures for each liquid at that temperature. Figures 4 and 5 shows the azeotropic effects which result in significantly decreased boiling points when two immiscible (insoluble) liquids such as benzene and water, or xylene and water are mixed.

Thus, in Figure 4 for the insoluble mixture of benzene and water, at 69° C (156° F) benzene's vapor pressure is 533 mm Hg, and water has a vapor pressure of 227 mm Hg. Since the total pressure is 760 mm, the mixture will boil at this

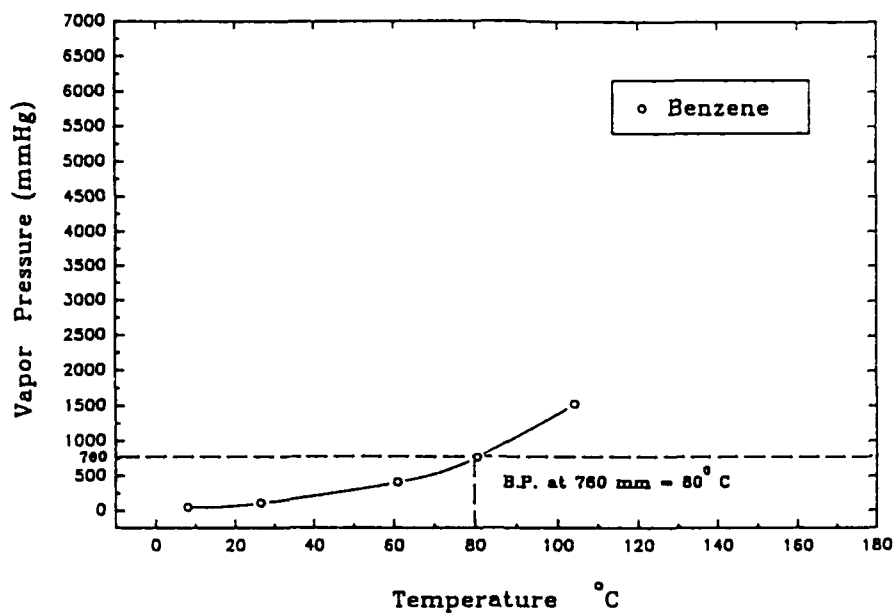


Figure 1.

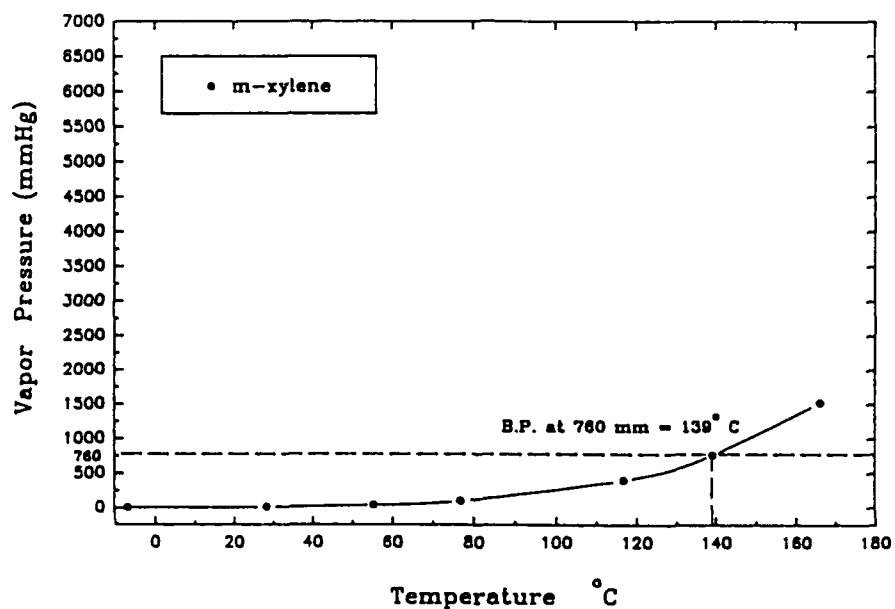


Figure 2.

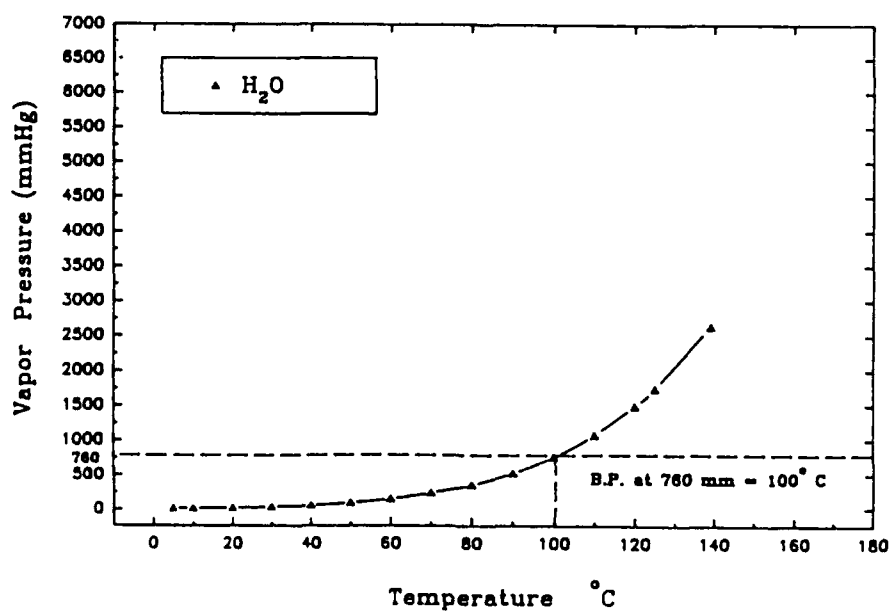


Figure 3.

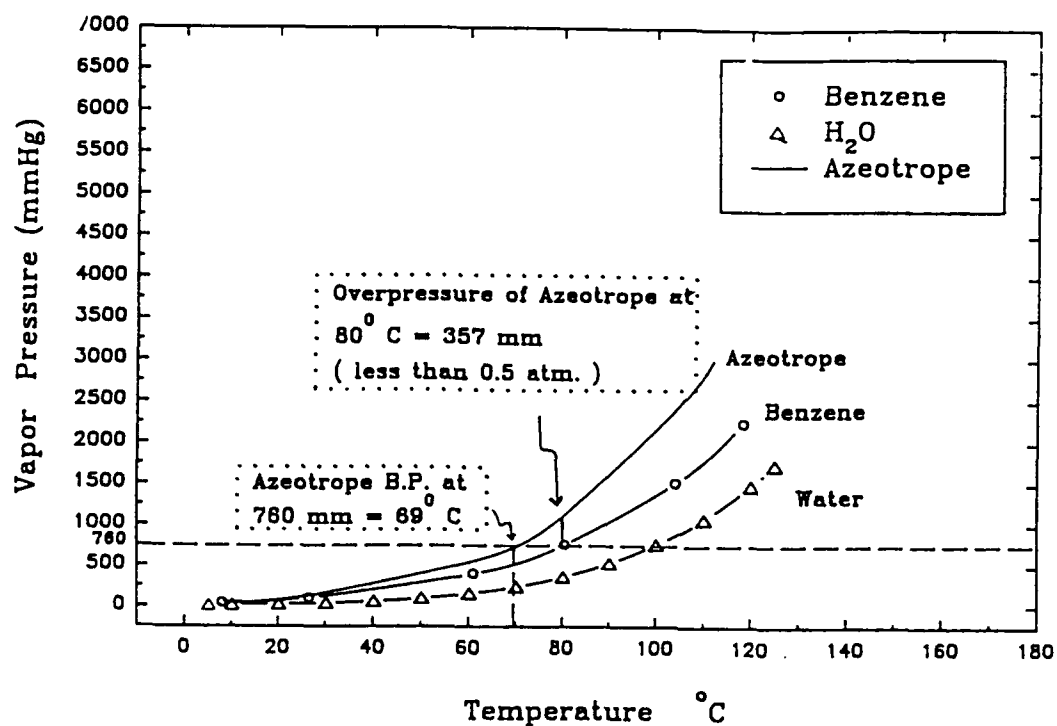


Figure 4.

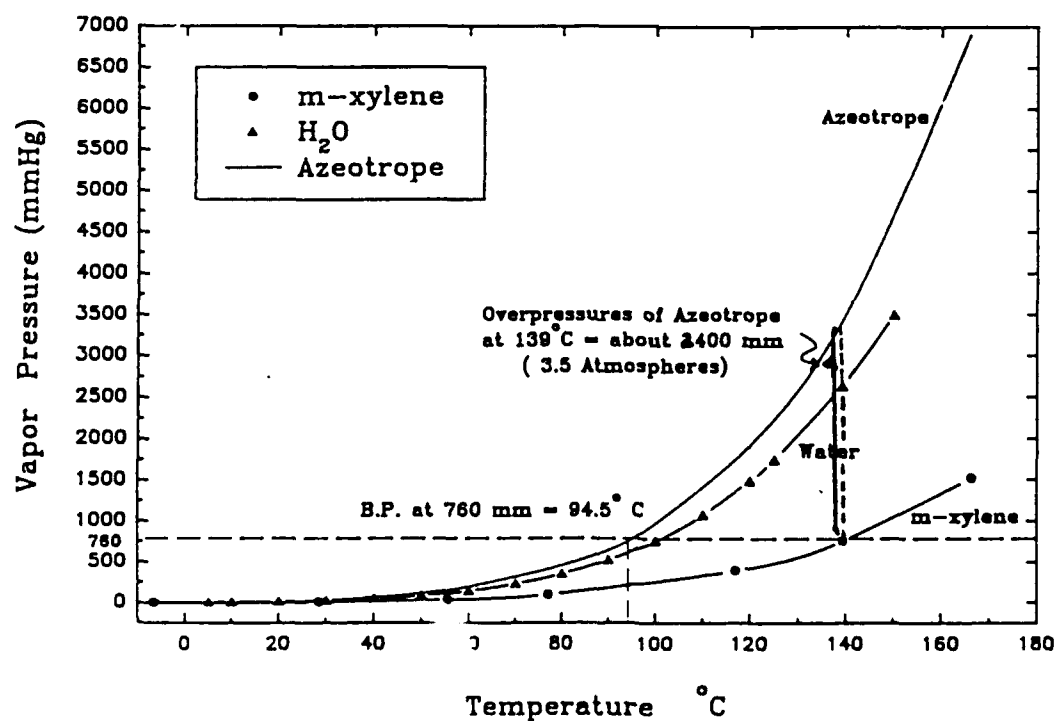


Figure 5.

lowered temperature, some 11° C (52° F) lower than the boiling point of pure benzene.

The effect becomes even more pronounced for higher boiling boint liquids, as can be seen in Figure 5 for the insoluble mixture of water and xylene. At 94.5° C (202° F), the vapor pressures of xylene and water total one atmosphere -- some 45° C (80° F) cooler than for xylene alone.

(Figures 4 and 5 also indicate another term, azeotropic overpressure; this important feature will be discussed in detail later in this report.)

All of these azeotropic considerations can have serious implications for the flammability of hydrocarbon fuels in contact with water -- e.,g., as in the case when fuel fires are being extinguished by water or water based extinguishing agents such as AFFF.

Fuel flammability, and intensity of fire for the fuel, is typically regarded in terms of the fuel's flash point -- i.e., temperature of the liquid at which its vapors are sufficiently present over the fuel to sustain a fire.⁵ The more flammable fuels are those with the lower flash points, and fuels with higher flash points are typically regarded as being more safe from the standpoint of such ignitions.

Due to these volatility considerations, aviation fuels have undergone dramatic changes since World War II. In 1951 the US Air Force and Army changed from a highly volatile blend of gasoline and kerosene to the less volatile JP-4 formulation still widely used by these services today. In 1952 the US Navy adopted a much less volatile blend (JP-5) as a result of the extreme fire fighting constraints peculiar to Navy carrier operations. In 1958 an intermediate blend (less volatile than JP-4, but more volatile than JP-5) was adopted as Jet A-1 fuel for use in commercial aviation; and since 1968 a slightly modified version of Jet A-1, designated as JP-8, has been gradually implemented for use by NATO and USAFE aircraft. The characteristics and compositions of these principal JP fuels are shown in Tables I and II.

TABLE I. FUEL CHARACTERISTICS⁶⁻¹²

US	JP-4	JP-5	JP-8	Jet A-1
UK	AVTAG		AVTUR	
NATO	F-40	F-44	F-34	F-35
Specific Gravity	0.77	0.83		0.80
Vapor Pressure, psi [RT]	3.0	LESS THAN 0.1 PSI		
Flash Point (Fahrenheit)	-20	+ 150		+ 125

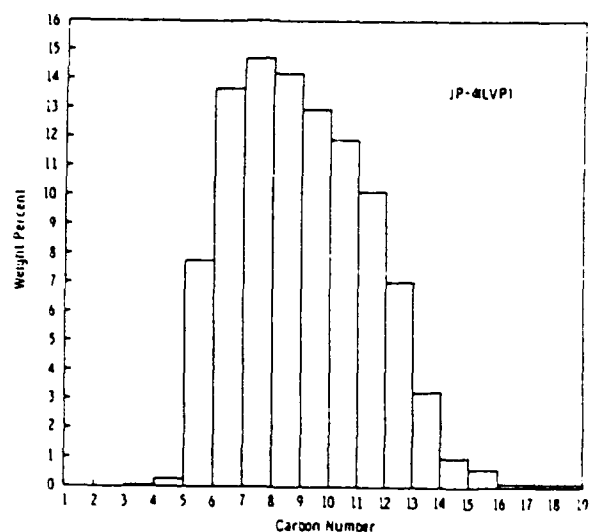
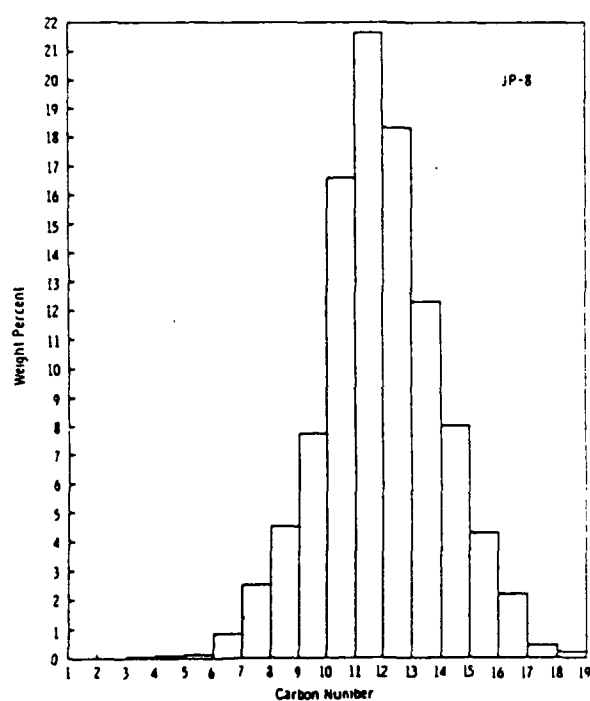
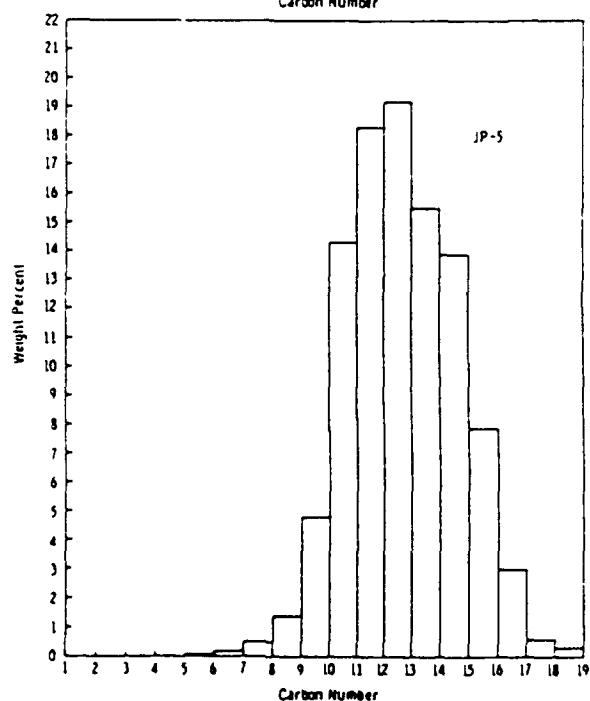


TABLE II. DISTRIBUTION OF
HYDROCARBONS IN JP-4, JP-5 AND
JP-8, DETERMINED BY SIMULATED
DISTILLATIONS⁶



In Table III are presented azeotropic and flash point data for C_6 through C_9 hydrocarbons -- these being prominent constituents of the relatively volatile JP-4 fuel widely used in the Air Force today. It can be noted:

TABLE III. AZEOTROPIC DATA FOR HYDROCARBON MIXTURES WITH WATER 13-15

Hydrocarbon	Formula	boiling point of pure hydrocarbon °F	boiling azeotropic mixture of hydrocarbon and water °F	decrease in hydrocarbon boiling point °F	per cent by volume, hydrocarbon in azeotrope	flash point, pure hydro- carbon °F [ref.16]	azeo- trope flash point °F
Hexane	C_6H_{14}	156	145	11	96	- 10	
Benzene	C_6H_6	176	156	20	92	12	
Cyclohexane	C_6H_{12}	177	156	21	93	- 1	U
Heptane	C_7H_{14}	208	174	34	91	30	N
Methyl- cyclohexane	C_7H_{12}	214	178	36	85	30	K
Toluene	C_7H_8	232	183	49	82	40	N
m-Xylene	C_8H_{10}	282	203	79	63	77	O
Nonane	C_9H_{20}	304	203	101	68	88	W
Cumene	C_9H_{12}	306	203	103	60	115	N

1. There is a significant increase in both boiling points and flash points of the pure hydrocarbons, with increasing numbers of carbons in the molecules (i.e., with increasing molecular weight).
2. Boiling points significantly decrease for each of these hydrocarbons when mixed with water.
3. There are bigger and bigger decreases in azeotropic boiling points of the hydrocarbon/water mixtures, with increasing numbers of carbons in these hydrocarbons. (For C hydrocarbons, reductions in boiling points for the hydrocarbon/water mixtures exceed 100° F, with most of the vapor comprised of the hydrocarbon fuel molecules!)

The matter of "azeotropic overpressures", referred to on page 7 in the discussion of Figures 4 and 5, was early regarded in this work as being an area of prime concern.

In a liquid fuel fire, the surface of the fire the burning liquid is at its boiling point. Although there will be a temperature gradient in the liquid fuel below the surface, there will be a significant fraction of the liquid fuel beneath the burning surface which will be at a considerably elevated temperature.

If a water-based extinguishing agent (e.g., fog, AFFF, or even a solid stream) is applied to this fire, the incoming water will also be significantly heated as it passes through the flame and into the burning liquid.

We will first examine a high volatility fuel such as benzene (see Figure 4), with a volatility representative of the major components of JP-4. If the benzene is heated to its boiling point (80°C , or 176°F), and water added at a rate such as to allow it to be heated to about this same temperature, the vapor pressures of the water (357 mm) and of benzene (760 mm) total now to 1117 mm. This is an overpressure of 357 mm (about $\frac{1}{2}$ atmosphere) in vapor pressure which has suddenly been installed in what had been a gently boiling liquid. The effect will be similar to that which we would observe if we heated the benzene to 92°C (198°F) in a closed pressure cooker, which would now show a pressure of about 7 psi or $\frac{1}{2}$ atmosphere on its dial. If we suddenly open the pressure cooker, the contents will erupt in very vigorous boiling. This same effect will be observed on addition of the water to the boiling benzene (or JP-4 type of fuel); and a somewhat (though perhaps not very greatly increased intensity in the fire will be observed.

Figure 5 illustrates a low volatility fuel such as xylene with a volatility representative of major components of JP-8. If xylene is heated to its boiling point (139°C , 282°F), and water added at a rate to allow it to be heated to about the same temperature, the vapor pressures of water (2,640 mm) and of benzene (760 mm) total now to 3,400 mm -- an overpressure of 2,640 mm or about $3\frac{1}{2}$ atmosphere in vapor pressure, again, suddenly unleashed in what had been a gently

boiling liquid. The effect is similar to that which we would observe if we heated the xylene to 200° C (about 400° F) in a closed pressure cooker, which would now show a pressure of about 52 psi or 3½ atmosphere on its dial. If we suddenly open the pressure cooker, the contents will erupt in very violent boiling. This same effect will be observed on addition of water to boiling xylene (or JP-8 type of fuel); and an extremely greatly pronounced increased intensity in the fire will be observed.

The questions then arose:

1. Since both boiling points and flash points of the pure hydrocarbons are both linked with the molecular weight of the fuel components; and since there is an obviously increased volatility of the hydrocarbon when water is present, is there a corresponding decrease in the flash points of these hydrocarbons, when mixed with water?
2. Is there a correspondingly increased reduction in flash points of higher molecular weight hydrocarbons as are found in JP-5 and JP-8 fuels, ordinarily considered to be less volatile and thus more fire safe than the more volatile JP-4 types of fuels?
3. What kinds of fire fighting situations would most contribute to unexpected increases of vaporization of the fuel, during extinguishing operations? (If there is an inordinate increase in such vaporization, it might be anticipated that there could be concomitant increases in flash back,

fireballing and similar unexpected flame flare-ups. Such situations could be particularly hazardous for large scale fire fighting operations.)

PROJECT OBJECTIVES

The overall objective of this project were to investigate the possible effect of water to cause jet fuels and similar hydrocarbon fuel compositions to burn faster and with greater intensities when the water is applied to the burning fuel, as a result of steam distillation effects.

In the first phase of work flash points were determined for representative volatile and non-volatile hydrocarbons, both in the dry state and when mixed with water.

In the second phase of the projects, effects on fire proclivities were studied, involving application of water to boiling hot hydrocarbon fuels of both high and low volatilities (as would be representative for JP-4 and JP-8 types of aviation fuels, respectively).

EXPERIMENTAL WORK. PART I. FLASH POINT DETERMINATIONS.

A HerzogTM semiautomatic electric Pensky-Martens closed cup electric flash point tester¹⁷ (designed for determination of flash points of mixtures of multi-phase systems, in accordance with ASTM D56 and D93 procedures), and a Herzog^M semiautomatic electric Cleveland open cup flash point

tester¹⁷ (designed for determination of flash points in accordance with ASTM D92 procedures) were used for:

1. Determining flash points of representative hydrocarbons of low and high molecular weight, of known flash point, by way of providing a standardization for the equipment.
2. Determine flash points of a variety of straight chain and branched chain alkanes, to ascertain possible structural effects which have not yet been examined with regard to flash points.
3. Determine flash points of variety of high and low molecular weight hydrocarbons mixed with water, to determine possible existence of steam distillation effects on flash points.
4. Attempt a quantification of such steam distillation effects, if any, for extrapolation to other hydrocarbon fuel components and mixtures.

EXPERIMENTAL RESULTS, PART I: FLASH POINT DETERMINATIONS FOR HYDROCARBONS WITH AND WITHOUT AZEOTROPIC WATER EFFECTS

The following observations were made:

1. Excellent agreement was achieved for all the many standard hydrocarbons using the open cup flash point tester, and for the few values available for closed cup determinations.
2. No reductions in flash points were observed for any of the standard hydrocarbons which were tested floating on water in the open cup tester.

As a corollary to this observation, it was also observed that water can be a very beneficial additive to immiscible (insoluble) fuels for flash point determinations, at least for fuels testing at flash points of 80° C or lower, and for the few with higher flash points, probably extending even higher than this value. Thus, heating is facilitated using the water. Nowhere nearly as much fuel is required -- only a thin film will suffice. The apparatus is much easier to clean after each determination. Accuracy is unimpaired -- the same readings are achieved with or without the water.

3. For flash point determinations for hydrocarbon on water, using the Pensky-Martens closed cup electric flash point tester, there were frequently no flash points observed. It was observed on such occasions that the build-up of water vapor served to occlude air from the closed cup, with the result that ignition was impaired or obviated.

EXPERIMENTAL RESULTS, PART II:

APPLICATION OF WATER TO BOILING HOT HYDROCARBON FUELS OF HIGH AND LOW VOLATILITIES (REPRESENTATIVE OF JP-4 AND JP-8)

For these determinations, an apparatus was constructed in accordance with the design indicated in Figure 6.

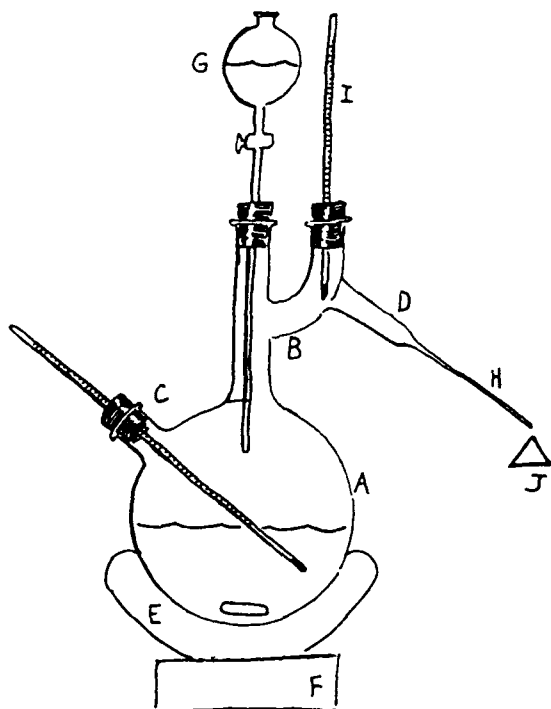


Figure 6. Ignition Flask

- A. 500-ml round bottom, with
- B. Two-necked Claisen head,
- C. Pot thermometer well,
- D. And side arm extending from Claisen head.
- E. Electric heating mantle.
- F. Magnetic stirrer.
- G. Addition funnel for adding water.
- H. Capillary extension tube from side arm tube.
- I. Head, pot temperature thermometers.
- J. Igniter

Not shown: Aluminum foil insulation around assembly; nitrogen tank for purging air from assembly; heating tape for side arm tube; emergency fire extinguishers

Procedures:

1. Add 100 ml fuel to flask, with magnetic stirring bar.
2. Purge air from assembly with nitrogen tank.
3. Set heating mantle and heating tape to about 20° C above boiling point of fuel.
4. When fuel begins to distill from capillary extension, light distillate with igniter.
5. Add 1 ml water from addition funnel, observe flame growth (if any) at capillary extension.

RESULTS OF EXPERIMENTS WITH XYLENE, AND XYLENE AND WATER;
AND WITH BENZENE, AND BENZENE AND WATER

1. For both xylene and benzene experiments, distillation rates were achieved (prior to adding water) which provided just enough fuel at the capillary to sustain a small flame at the capillary extension tube.
2. For distilling benzene, addition of water did not materially increase the magnitude of the flame.
3. For distilling xylene, addition of water resulted in a huge increase in flame size; see Figures 7 and 8 below.



Figure 7. Xylene flame
prior to
adding water.



Figure 8. Xylene flame is
greatly intensified
by adding of water.

RAMIFICATIONS OF WATER AZEOTROPING EFFECTS FOR JP-8, JP-5
AND JET A-1 (AND SIMILAR LOW VOLATILITY HIGH BOILING
POINT FUEL) FIREFIGHTING CONSIDERATIONS

1. Operational firefighting

No previous attention appears to have been directed to the possibility of increased flammability hazards arising from azeotroping effects from application of water systems to hydrocarbon fuel fires. Statements to the contrary have been encountered in responsible fire manuals:

" ... water ... entrained in fuel ... is not particularly significant from a fire hazard viewpoint" ¹⁸ This is valid for firefighting implications for the more volatile JP-4 type fuels; but it is an over-simplification from the standpoint of highly possible increases in volatilities for less volatile fuel blends such as JP-8, JP-5 and Jet-A1.

There is at least one instance in the fire fighting literature which can now possibly be reinterpreted in the light of a possible water/fuel azeotroping effect.

On 26 May 1981 an EA-6B crashed into several F-14's while landing on the US Navy carrier NIMITZ (CVAN 68). In the ensuing fire 14 men were killed and 42 injured, with \$60 million damages to the carrier and its planes. Fire fighting efforts commenced immediately, using water hoses and AFFF washdown systems (although the AFFF systems were not deployed until well into the fire fighting effort). In a subsequent

investigation it was suggested that possibly there had been contamination of JP-5 fuel in the Navy aircraft by JP-4 fuel as a result of refuelling from an Air Force tanker; and that there had been a reduction in flash point of the Navy jet fuel as a result of the possible admixture with the more volatile JP-4.¹⁹

A possibility that greatly increased volatilization occurred when the water based extinguishing agents (fog or AFFF) contacted the hot fuel, should be investigated from the standpoint of future fire fighting technologies. (It should be noted that Halon extinguishing agents may be unavailable in the future, with an increased reliance on water based extinguishing systems. From the standpoint of Air Force interests, if there is a conversion from more volatile JP-4 fuel stocks to less volatile JP-8 fuel; and if JP-8 actually turns out to be prone to unanticipated increased vaporization rates in the presence of water, due to azeotropic effects, the need for an in depth evaluation of this effect would assume even greater dimensions of urgency. It should also be noted that the Navy is now using low-volatility JP-5 fuel, and that commercial aircraft are now exclusively fueled with low volatility Jet A-1 [essentially identical to JP-8]. Thus, need can be established for examination of the azeotroping effects from the standpoint of Navy and commercial aviation interests, as well.]

In summary, the following implications pertain for azeotropic water effects in operational firefighting considerations:

- (1) Application of water onto burning fuels insoluble in water will result in an increase rate of volatilization of the fuel, and a correspondingly increased fire intensity will result.
- (2) The effect is particularly pronounced for less volatile "fire-safe" fuels such as JP-8, JP-5 and Jet A-1. (See Figures 9 and 10.)
- (3) Due to unexpected high increases in rates of volatilization which can result with low volatility fuels on application of AAAF, water fog or other water-based firefighting agent, it may be best to use halon or alternative halons for supplementary extinguishment.
- (4) A need exists for increased firefighter awareness of unexpectedly high increases in rates of volatilization for low volatility fuel fires, when using water-based extinguishing agents.
- (5) Water suspended in the fuel before the fire will not materially affect the flash point.



Figure 9. Water used in extinguishment of low boiling point (high volatility) fuel fires: no anomalous effect. (E.g., JP-4 fires.)



Figure 10. Water used in extinguishment of high boiling point (low volatility) fuel fires: increased fire intensities! (E.g., JP-8, JP-5, Jet A-1 fires.)

2. Firefighting Training Considerations

In typical firefighting training exercises, a large fire pit is partially filled with water to provide a flat surface for fuel which is then layered to a depth of an inch or so over the water (the flat water surface minimizes fuel volume requirements). JP-4 is typically used for CONUS Air Force training requirements; JP-8 for USAFE training; Jet A-1 for commercial aviation training requirements, and JP-5 for Navy training). A diagram of a typical fire pit assembly is shown in Figure 11.

As can be seen in Figure 11, there is therefore a very sharp temperature gradient in the very thin layer of burning liquid fuel. At the utmost surface, the fuel temperature will be at its boiling point (i.e., 139° C or 282° F for xylene); but an inch or so below this, at the interface of the fuel layer with the underlying water, the temperature will have dropped to the ambient water temperature (typically no more than 90° F). Therefore, almost all of the fuel will be at a temperature which is far below its azeotropic boiling point (in the case of xylene, 203° F).

Thus, for even the most non-volatile hydrocarbon fuel such as JP-5, there will be no observed increase in rate of volatilization of the burning fuel when water-based extinguishing agents are applied to the fire!

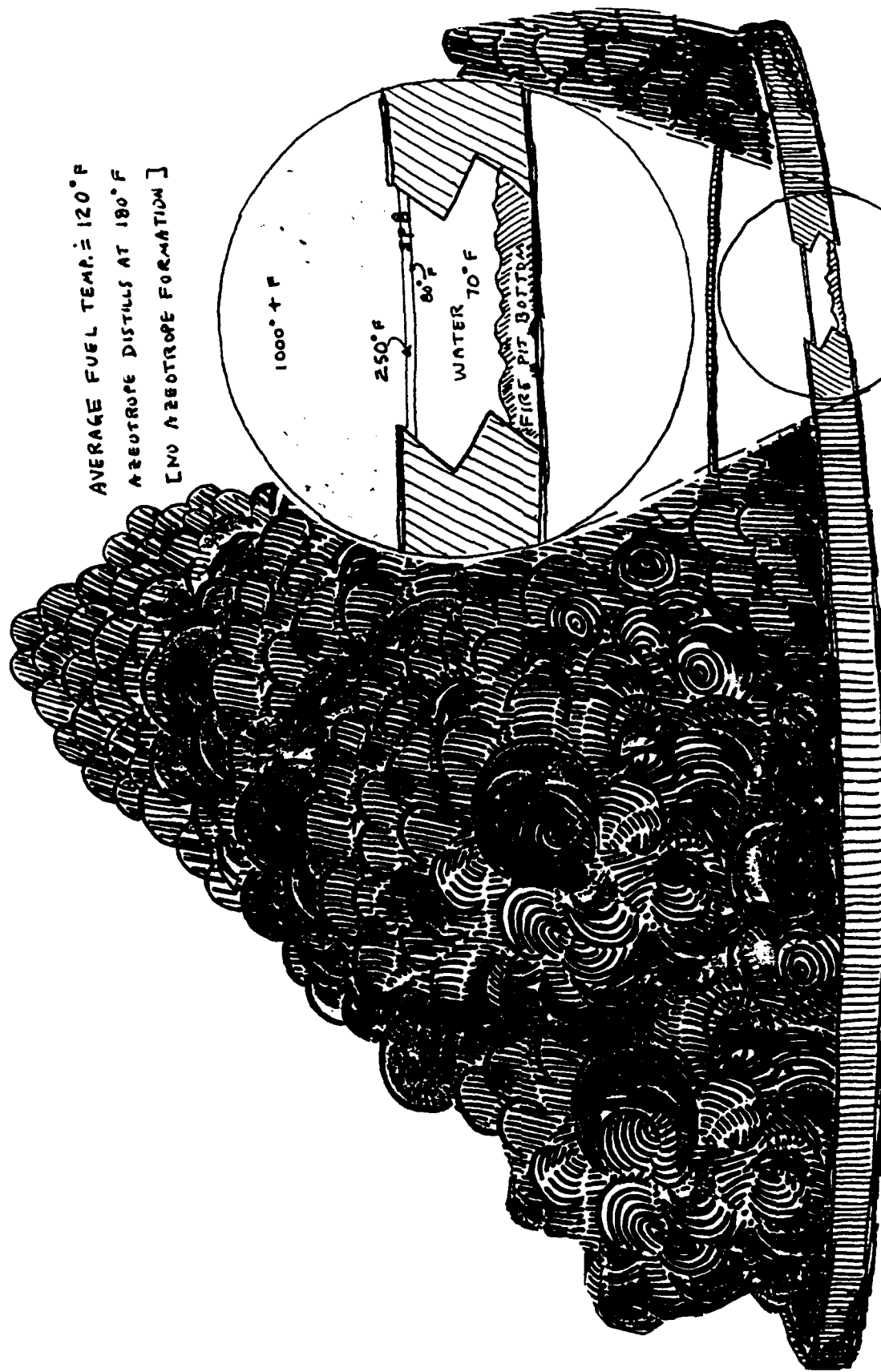


Figure 11. Typical firetraining fire pit assembly. (No significant increase in rate of volatization will be observed for fuel floating on a water surface.)

It is therefore recommended that firefighting tests be conducted with water fog and AFFF on JP-8 fires in a 100-foot fire put with no water layer present. The pit should contain three-dimensional metal structures above the fuel level, to provide real-life hot spots for heating incoming water.

It is also recommended that existing film and video coverage on JP-8/JP-5/Jet A-1 fires be studied to ascertain the extent of azeotropic steam distillation effects which may have been experienced in previous commercial and military aircraft or POL fires.

REFERENCES

(All references are unclassified.)

1. Hucknall, D. J. "Chemistry of Hydrocarbon Combustion", New York, Chapman and Hall, 1957.
2. Verhvalin, C. H. "Fire Protection Manual", 3rd ed. (Gulf Publishing Co., Houston), Vol. 1, pp. 139-143.
3. Hunt, H. "Physical Chemistry", New York, Thomas Y. Crowell Co., 1947; pp. 233-234.
4. Prutton, C. F.; Maron, S. H. "Fundamental Principles of Physical Chemistry" (MacMillan Co., New York); p. 177 (1951).
5. "Fire Hazard Properties, Flammable Liquids, Gases and Volatile Solids", Nat'l Fire Protection Assn., Boston, MA, 1960.
6. Parts, L.; Bucher, T. J.; Botteri, P.; Cretcher, E. "Integral Aircraft Fuel Tank Classification"; AFAPL-TR-79-2092, Aero Propuls. Labs, W-P AFB, OH (1985).
7. Richardson, S. A. "JP-8, A Potential Replacement for JP-4 Jet Fuel"; Air Univ., Maxwell AFB, AL (May 1973).
8. Beery, G. T., et al. "Assessment of JP8 as a Replacement for the Air Force Standard Jet Fuel JP-4". AFAPL-TR-74-71, Part I; Aero Propulsion Lab, WP-AFB, OH (1975).
9. Botteri, B. P. "Flammability Properties of Jet Fuels and Techniques for Fire Explosion Suppression"; AGARD Conferences Proceedings, NATO, pp. 13-1 ro 13-11.
10. Kutcha, J. M.; Clodfelter, R. G. "Aircraft Mishap Fire Pattern Investigations"; AFWAL-TR-85-2057, Aero Propulsion Labs, W-P AFB, OH (1985).
11. Frame, E. A. "Behavior of Fuels at Low Temperatures"; Interim Report, AFLRL 138; US Army Fuels and Lubricants Res. Lab., Southwest Res. Inst., San Antonio, TX (1980).
12. Gardner, L.; Whyte, R. "Jet Fuel Specifications". AGARD Conference Proceedings No. 84 on Aircraft Fuels, Lubricants and Fire Safety, NATO (1984).
13. Lange, N. A. "Handbook of Chemistry", 9th ed., Sandusky, OH, Handbook Publishers, Inc., 1956; p 1485.

14. Claxton, G. "Physical and Azeotropic Data" (Nat'l Benzole and Allied Products Assn., Cambridge, England); pp. 104 - 132.
15. Weast, R. C. "Handbook of Chemistry and Physics", 56th ed. Boca Raton, FL, CRC Press, 1975; pp. D1 - D36.
16. McCracken, D. J. "Hydrocarbon Combustion and Physical Properties", BRL-1496; Ballistic Res. Labs, Aberdeen Proving Ground, MD (1970).
17. VWR Scientific Co. Catalog, 1989-1990; p. 1044.
18. Appendix A, Standard 407, NFPA Fire Codes (National Fire Prevention Association, Boston, MA); p. 407-23.
19. Carhart, H. W., et al. "Aircraft Carrier Flight Deck Fire Fighting Tactics and Equipment Evaluation Tests". NRL Memorandum Report 5952, Feb. 1987.

ACKNOWLEDGMENTS

This research was sponsored and supported by the US Air Force Systems Command and the Air Force Office of Scientific Research. Support was also extended by the College of Pure and Applied Science of the University of Lowell (Massachusetts), and the Massachusetts State Fire Commission, (Boston). Grateful acknowledgement is extended to Universal Energy Systems, Inc. for administration of work under the aegis of the Air Force Summer Faculty Research Program. We are also grateful for valuable information, advice and suggestions provided by Mr. Richard Vickers, Chief, Fire Protection and Crash Rescue Systems Branch [RDCF] at Tyndall AFB, Florida, and Mr. Charles Risinger, my project officer at RDCF, and Mr. Andrew Poulis, Chief Librarian of the Air Force Engineering and Services Center at Tyndall AFB; Dr. George Geyer of the Federal Aviation Administration's Technical Center, Atlantic City; Assistant Chief Mark S. Lawlor, Hickham AFB, Hawaii; Professors James Pierce of the University of Lowell and Paul L. Damour of St. Anselm College, Manchester, NH; and Stephen Bistany, Michael Orroth, Iurie Schwartz, Gregory Searle, and Kevin White of my research team at the University of Lowell.

1990 USAF-UES RESEARCH INITIATION GRANT

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Bolling AFB, DC

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

Prepared by:	Emerson L. Besch, PhD
Academic Rank:	Professor
Department and University	Department of Physiological Sciences University of Florida
Research Location:	Gainesville, FL 32611-0633
Date:	7 December 1990
Contract No.	F49620-88-C-0053/SB5881-0378

EFFECT OF SIMULATED JET AIRCRAFT NOISE ON DOMESTIC GOATS

by

Emerson L. Besch, PhD

ABSTRACT

Physiological and behavioral adjustments to environmental stressors can provide information on the well-being of animals. In a study of the physiological effects of jet aircraft noise, adult, mixed breed, domestic male and female goats were allowed to adjust to pens and blood sampling procedures prior to measuring the stressor effects on cardiac frequency, plasma cortisol, and white blood cell counts. The single-event and multiple-event stimuli were provided by a noise simulation system that exposed animals to noise intensities that nominally would be encountered during jet aircraft operations. Goats were chosen as animal models because they are easily managed and maintained, readily available, comparatively inexpensive, ruminants, and ideally suited as a wildlife analog. As a result of the study reported here, the usefulness of a locally fabricated radio transmitter as a non-invasive technique to evaluate heart rates of domestic goats was successfully demonstrated through comparison with data obtained from commercially available heart rate monitors. Although behavioral responses to noise stimuli were observed, heart rate responses revealed no information that was not apparent from changes in behavior. Heart rates were increased 55-93% in noise-exposed compared to control goats. Also, an indirect, but non-significant, correlation between the noise exposure interval and net percent increase in heart rate was observed. Attempts to identify heart rate as a predictor of plasma cortisol were unsuccessful. Also, no statistically significant changes in plasma cortisol, lymphocytes, or neutrophils were detected in goats following exposure to single-event or multiple-event noise intensities. Nonetheless, an overall assessment of the results suggests that additional studies on the effects of noise on domestic goats are needed.

Acknowledgements

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for funding and the Universal Energy Systems, Incorporated, for administering the 1990 Research Initiation Program. This funding provided me an opportunity to conduct follow-up studies on the research trials completed at Tyndall AFB, Florida, in the Summer of 1989 as part of the USAF-UES Summer Faculty Research Program.

In completing the objectives of this program at the University of Florida, many individuals played a role. In particular, I would like to thank Robin L. Brigmon, Biological Scientist, and Stephen M. Varosi and Dawn L. Miller, Student Assistants, all of whom work in my laboratory and all of whom made significant contributions in collecting, recording, and analyzing data. Special thanks are offered to Edwin L. Crews, Agricultural Technician III, of the IFAS Horse Teaching Unit, for his "can do" attitude and general assistance regarding animal care and management and to Jamiu Williams, a High School Summer Research Apprentice, who seemed to thrive on being offered new experiences and challenges. Others who were especially helpful include Dr. Christophe W. Lombard, Associate Professor of Cardiology, College of Veterinary Medicine, who loaned the electrocardiorecorders and provided assistance in analyzing heart rate data; Major Michael M. Thompson, USAF, Chief, Flight Environmental Impact Program, Tyndall AFB, FL, for technical advice and for arranging the loan of the noise simulation system; Major Robert Kull, USAF, Program Manager, Noise and Sonic Boom Impact Technology (NSBIT) Program, Wright-Patterson AFB, OH, for technical advice; and, Dr. Joan Scott, Environics Division, USAF Engineering and Services Center, Tyndall AFB, FL, for providing the stimulus that led to the initiation of the "Goat Project" in the first place. And, last but not least, thanks also are given to the USAFNSBIT Technical Program, Wright-Patterson AFB, for the loan of the Noise Simulation System.

I. INTRODUCTION:

Since the introduction of jet aircraft into its inventory, the U. S. Air Force (USAF) has had to deal with complaints from the public regarding subsonic and supersonic noise. Public concern over noise pollution has been exemplified by campaigns to restrict flights of military aircraft at supersonic speeds over areas of high human population (Ewbank, 1977). The result is that many of those flights now are over the sea or other areas of low human density (e.g., National Parks, wildlife conservation areas).

In the 1960-70s, the focus of many noise studies was on occasional booms which did not appear to have much effect on an animal's behavior, except for "startle" response (Bell, 1972). All species studied (e.g., cattle, sheep, horses, pigs, poultry) appeared to physiologically adjust to sonic booms (Casady and Lehman, 1967; Espmark *et al.*, 1974; Ewbank, 1977; Shotton, 1982) and subsonic flight (Cotterau, 1978) with no loss of productivity. In general, behavioral responses were minimal except for the avian species. For the latter, animals "scatter" or "crowd" (Ewbank, 1977) or tend to exhibit an "alert" reaction and "movement" away from the source of the sound (Bond, 1971). Sonic booms cause a "startle" response in cattle and sheep but this decreases as the duration of the exposure increases; sheep appear to be more disturbed by sonic booms when they are standing compared to lying (Espmark *et al.*, 1974). Little information is available on the effects of noise from subsonic flights or sonic booms on goats.

Research trials conducted at Tyndall AFB, Florida, in the Summer of 1989 assessed the response of domestic goats to noise from low-flying jet aircraft. These field studies were part of the UES Summer Faculty Research Program (SFRP)/Graduate Student Research Program (GSRP). The Final Reports of the Summer Fellow (Besch, 1989) and Graduate Student Researcher (Zern, 1989) described the difficulties associated with conducting field research. In particular, it was very difficult to control variables (e.g., weather, noise intensity and duration, feed and water intake) which can modify animal function and response to the noise stressor. While there is evidence to suggest that some of the goats were physiologically affected by the jet aircraft noise, others displayed responses that appeared to be unrelated to the noise stressor.

Although there were no significant differences between the baseline heart rates between the two groups of TAFB goats, the goats located near the jet aircraft runway

(Group 1) displayed a decreasing heart rate over time which was consistent with their behavior. The control group (Group 2) of goats whose pens were in a semi-isolated location on the base, maintained an elevated heart rate and never seemed to adjust to their new environment, displayed apprehensiveness, proved difficult to manage, and exhibited an uneasiness at the sight of people. Nevertheless, only the Group 1 goats were exposed to elevated noise levels associated with takeoffs and landings of jet aircraft, but both groups displayed hematological changes (e.g., relative lymphopenia and neutrophilia) that are considered to be adrenocortical-mediated stressor responses.

From the above observations it is difficult to separate the physiological and behavioral responses due to the jet aircraft noise stressor from the responses due to other factors such as difference in site location (e.g., Group 2 animals may have been too isolated from humans or too close to the track or trail of predator animals such as a coyote). It also is possible that the observed changes in the goats represent a summation of responses due to simultaneous exposure to multiple stressors (e.g., noise, isolation, new surroundings). One way to deal with the latter is to perform experiments under laboratory conditions where all variables--including the experimental--can be properly controlled. In this way, the relationship between one experimental variable (i.e., noise intensity) and the goat's response (i.e., physiological strain) could be separated from other potential stressor responses. This was the basis for the proposal that led to the research reported here.

II. OBJECTIVES

A. GOAL OF RESEARCH

The primary goal of this research was to determine the effects of simulated jet aircraft noise on the physiological well-being of domestic goats. The goat was chosen as the animal model because it is

- A. An ubiquitous domestic animal.
- B. Easily managed and maintained.
- C. Readily available.
- D. Comparatively inexpensive.
- E. A ruminant.
- F. Ideally suited as a wildlife analog.

B. SPECIFIC OBJECTIVES

To attain the research goal, the experimental plan was subdivided into four specific research objectives as follows:

- A. VALIDATION OF A HEART RATE RADIO TRANSMITTER
- B. HEART RATE AS A PREDICTOR OF PLASMA CORTISOL LEVELS
- C. BEHAVIORAL RESPONSES OF GOATS TO SIMULATED JET AIRCRAFT NOISE
- D. PHYSIOLOGICAL RESPONSES OF GOATS TO SIMULATED JET AIRCRAFT NOISE

C. ANIMAL CARE AND USE APPROVAL

The goats used in the research described here were obtained from and health care was provided by the University of Florida, Health Center Animal Resources Unit (HCARU). Animal care and use procedures were in accordance with national standards as described in the Guide for the Care and Use of Laboratory Animals (ILAR, 1985) and Guide for the Care and Use of Agricultural Animals in Agricultural Research and Teaching, (Anonymous, 1988). The research plan and protocols were reviewed and approved by the University of Florida Animal Care and Use Committee prior to initiation of animal experimentation.

III. OBJECTIVE NO. 1: VALIDATION OF A HEART RATE RADIO TRANSMITTER

A. INTRODUCTION

Heart rate has been used as an indicator of stressor response because it is correlated with energy expenditures (Richards and Lawrence, 1984) and is a sensitive measure of arousal (MacArthur *et al.*, 1982). It also has been reported (Harlow *et al.*, 1987) that remote monitoring of cardiac frequency can be used as a predictor of adrenal function and, therefore, the potential immunologic condition of an animal during exposure to an environmental stressor.

But, commercially available radiotelemetry and electronic devices for measuring heart rate are expensive. Further, radiotelemetry requires a signal receiver unique to the transmitter and duration of use is limited by the operating life of the transmitter's

battery while electronic devices are limited by the operational life of the audio cassette on which heart rate is recorded. Also, radio transmitters require surgical implantation which exposes the animal to some trauma and necessitates recovery time.

These objections could be overcome with a reliable, extended-use, and externally mounted radiotelemetry device for measuring heart rates in unrestrained animals. A radio transmitter which fulfilled these requirements was designed, fabricated, and used to measure heart rates in unrestrained domestic goats. Heart rate data were compared to those simultaneously obtained from the same animal using a commercially available and externally mounted electronic heart rate monitor. The results of those comparisons, together with a description of the heart rate transmitter, are reported here.

B. MATERIALS AND METHODS

Animals: Four mixed-breed male and female goats with a body mass of 22.4 ± 1.0 kg (Mean \pm SE) and between 6 and 13 months of age were used in this study. Of these, 1 male and 1 female came from one subgroup of 6 goats and 1 male and 1 female came from another subgroup of 7 goats.

Animal environment: Each subgroup of goats was housed in a separate enclosed pen made from converted horse stalls which measured about 3.7m in length and 7.6m in width. Thus, each goat was provided at least 4.7m^2 of pen space. The pens contained cathedral ceilings and were ventilated with outside air through 3 half-doors (upper half of door always remained open) and thermostatically-controlled window-mounted exhaust fans provided air movement. Air temperature in each pen was monitored with thermistors (Omega, Type T, Stamford, CT) placed in three different locations, 1.5m above the floor. Air temperature from each thermistor was recorded using a data acquisition system (Model 2200B, John Fluke Mfg Co., Everett, WA) every 60 minutes. There were no statistically significant ($P > 0.05$) differences between goat pen temperatures which cycled daily between 21.1 ± 0.9 (Mean \pm SE) and $33.0 \pm 0.5^\circ\text{C}$ (Control) and 22.2 ± 0.9 and $32.5 \pm 0.6^\circ\text{C}$ (Experimental) during each 3.5 wk data collection period.

Feed and Water: Each group of goats was fed (0.34kg/goat/da) goat chow (Purina Mills, Inc, Goat Chow[®]) and hay each morning. All goat chow was weighed on a temperature compensated spring scale (Chatillon, Type 027). Water was changed each

morning (7:30-8:30 am) and evening (5:30-6:30 pm) and was provided to the goats ad libitum.

Telemetry System: The heart rate telemetry system (Figure III.1) was a modification of the one used to measure body temperature (Varosi, et al., 1990). The modifications included the addition of a lowpass filter (LPF), a one-shot trigger circuit, and a newly designed serial interface. The RF signal from the heart rate transmitter was received by an antenna and amplified, as necessary, for long coaxial cable transmissions. The antenna was a RG59/U female connector with a 1-2m center wire and two 1-2m ground plane wires soldered to it. The signal from the antenna was fed into a programmable receiver/scanner (PRO-2004, Radio Shack, Ft. Worth, TX). The receiver/scanner had pre-programmed channels for the unique frequencies of the individual radio transmitters. The audio output of the receiver/scanner was fed to a lowpass filter and one-shot trigger circuit which produced a pulse waveform that can be counted by a frequency counter (Model 1910A, John Fluke Mfg Co., Everett, WA). A serial interface (fabricated in the laboratory) converted the digital output of the frequency counter to a RS-232 serial port for interfacing with an IBM-compatible personal computer (PS/2 Model 50, IBM, Boca Raton, FL) where it was recorded each minute and stored in ASCII format for further analysis. Software, written in BASIC, controlled data acquisition, data storage, and sequencing of the receiver/scanner to the next programmed channel.

Heart Rate Monitors and Transmitters: Heart rate simultaneously was measured in each goat for a period of at least 2 hr using a battery-operated electrocardiocorder[®] cassette recorder (Model 456B, Del Mar Avionics, Irvine, CA) and a heart rate FM transmitter. The electrocardiocorder[®] contained an audio tape cassette (Holter 24-hr cassettes, Del Mar Avionics, Cincinnati, OH) on which the heart rate was recorded, each minute, for subsequent analysis by computer (Holter Analysis System, IBM PC Model 152, Del Mar Avionics, Irvine, CA). The FM radio transmitter (Figure III.2) was similar to the one used to measure body temperature except that the temperature sensitive oscillator has been replaced with a high gain electrocardiograph (ECG) amplifier requiring a 4-volt, center-tapped power supply (BT1, 2, 3). A remote-controlled power switch also was included but, because of the center-tapped power supply, only the transmitter was turned "on" and "off"; the ECG amplifier section was left "on" continuously with a quiescent current drain of only 0.01 mA. By incorporating a MOSFET (Varosi et al., 1989), both the transmitter and amplifier sections could be switched "on" and "off" but this would only complicate fabrication. The ECG amplifier was constructed with a CMOS

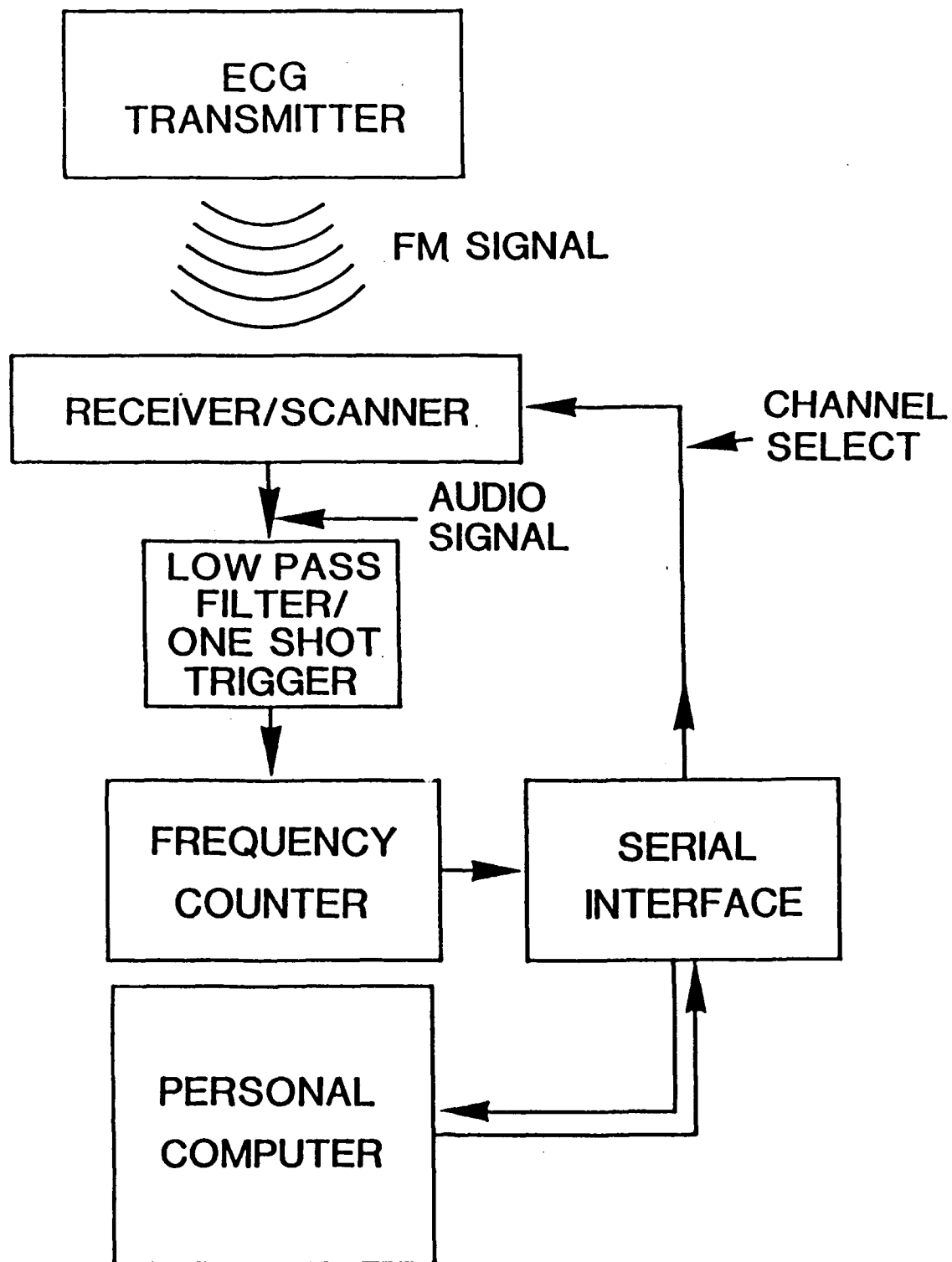
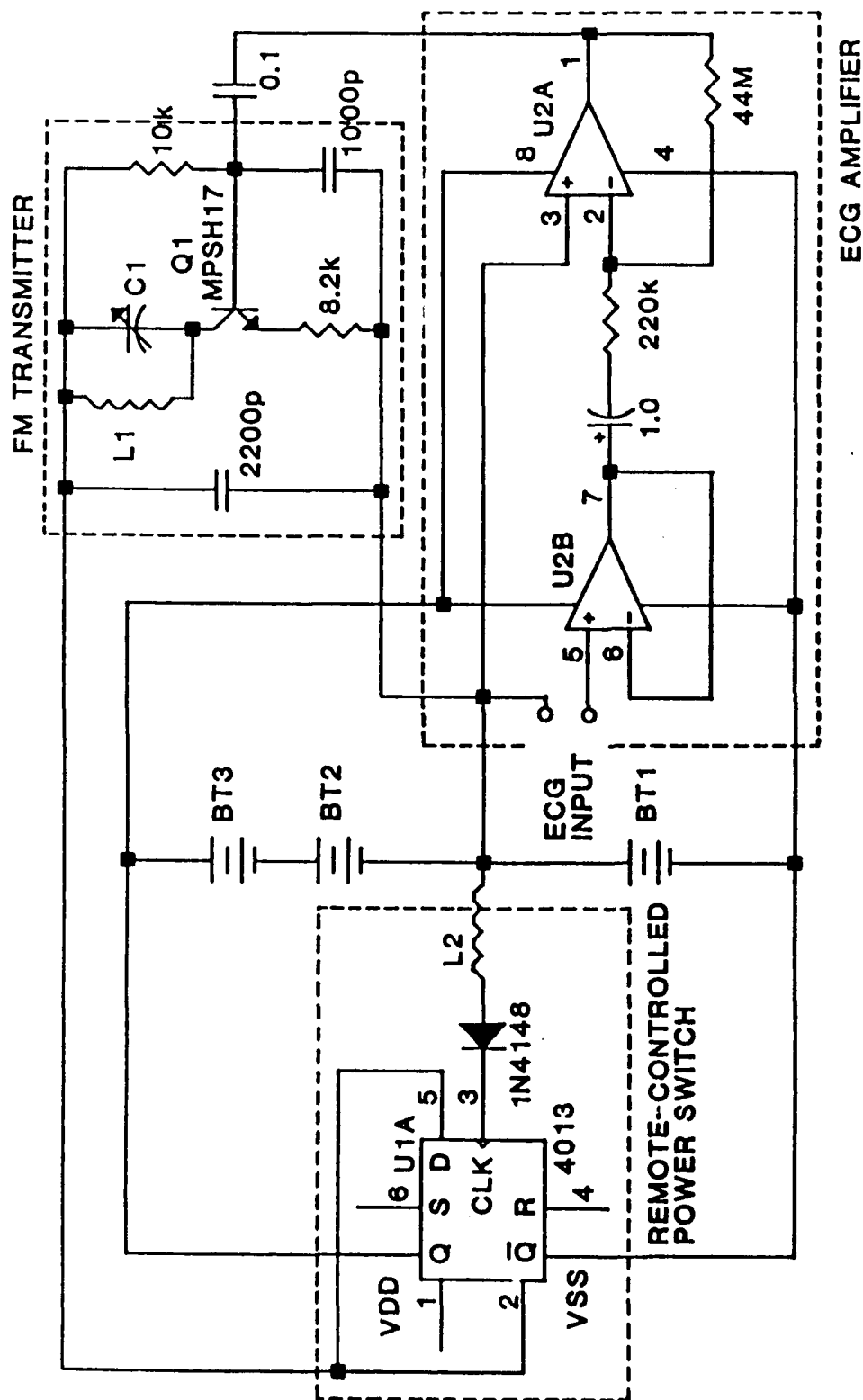


Figure III.1.

BASIC COMPONENTS OF TELEMETRY SYSTEM.



L1--: TURNS 22 AWG, 1CM DIA.
 L2-SWITCH TRIGGER COIL (200 TURNS 36 AWG, 1.3CM DIA.)
 BT1,2,3--EP675 MERCURY 1.35V
 C1--3-20 pF TRIMMER (MOUSER ME242-2820)
 U1--MC14013BCP
 U2--TLC27L2BCP

Figure III.2. ECG TRANSMITTER

op-amp which had an extremely high input impedance and very low power drain. The first stage, U2A, was a high impedance unity gain buffer which prevents skin and electrode resistance from attenuating the ECG signal before amplification. The output of U2A was amplified by U2B with a gain of 440. The output of U2A modulated the FM transmitter with the ECG profile detected at the ECG input. The FM transmitter carrier frequency was tunable, via trimmer capacitor C1, between 88-170 MHz.

Lowpass Filter: The scanner audio output contained high frequency noise as well as the desired low frequency ECG signal. To extract the ECG signal, a three-stage lowpass filter (Figure III.3) using OP-amps U1A, U2A, and U3A was employed. Each identical stage was a second-order, unity DC gain, active lowpass filter with a -3dB cutoff at 48Hz. The output of the third stage was amplified by U1B with adjustable gain of 1 to 21. The output of U1B was the filtered ECG signal. At this point, an oscilloscope was used to view the signal to ensure that the highest peak of the signal was positive, which should be the case provided the leads are properly placed.

One-Shot Trigger: The ECG signal was compared by U4B to a level preset by a 10k potentiometer (Figure III.3) to reduce the possibility of muscle artifact and other noise being interpreted as an ECG signal. The ECG signal usually was much larger than the other signals and setting the comparator level to 75% of the highest positive ECG peak was sufficient to reduce false triggers. To further reduce the incidence of false triggers, a one-shot trigger circuit U3A and U3B was introduced. When triggered, the output U3B went high (+8V) and stayed high for a period set to about 50-75% of the ECG signal period (adjusted by a 1M potentiometer) during which the circuit was immune to further triggering. This circuit reduced the possibility of false triggers between ECG pulses from noise and secondary peaks of strong ECG signals. The one-shot trigger output was sent to the frequency counter which counted positive rising edges of the pulse waveform which corresponded to ECG pulses.

External mounting of heart rate monitors and radio transmitters: After shaving an approximate 13cm X 13cm area on each side of a goat's thorax, both areas were thoroughly cleaned with 70% alcohol saturated using 4cm X 4cm gauze pads. Before recording heart rate, flexible elastic bandaging tape (Flexo®, Horse Health Products, Inc., Aiken, SC) was wrapped around the animal to protect the heart rate monitor, FM transmitter, electrodes, and electrode leads from dislocation and from the other goats. At the end of each sampling period, the monitor, radio transmitter, and electrodes were

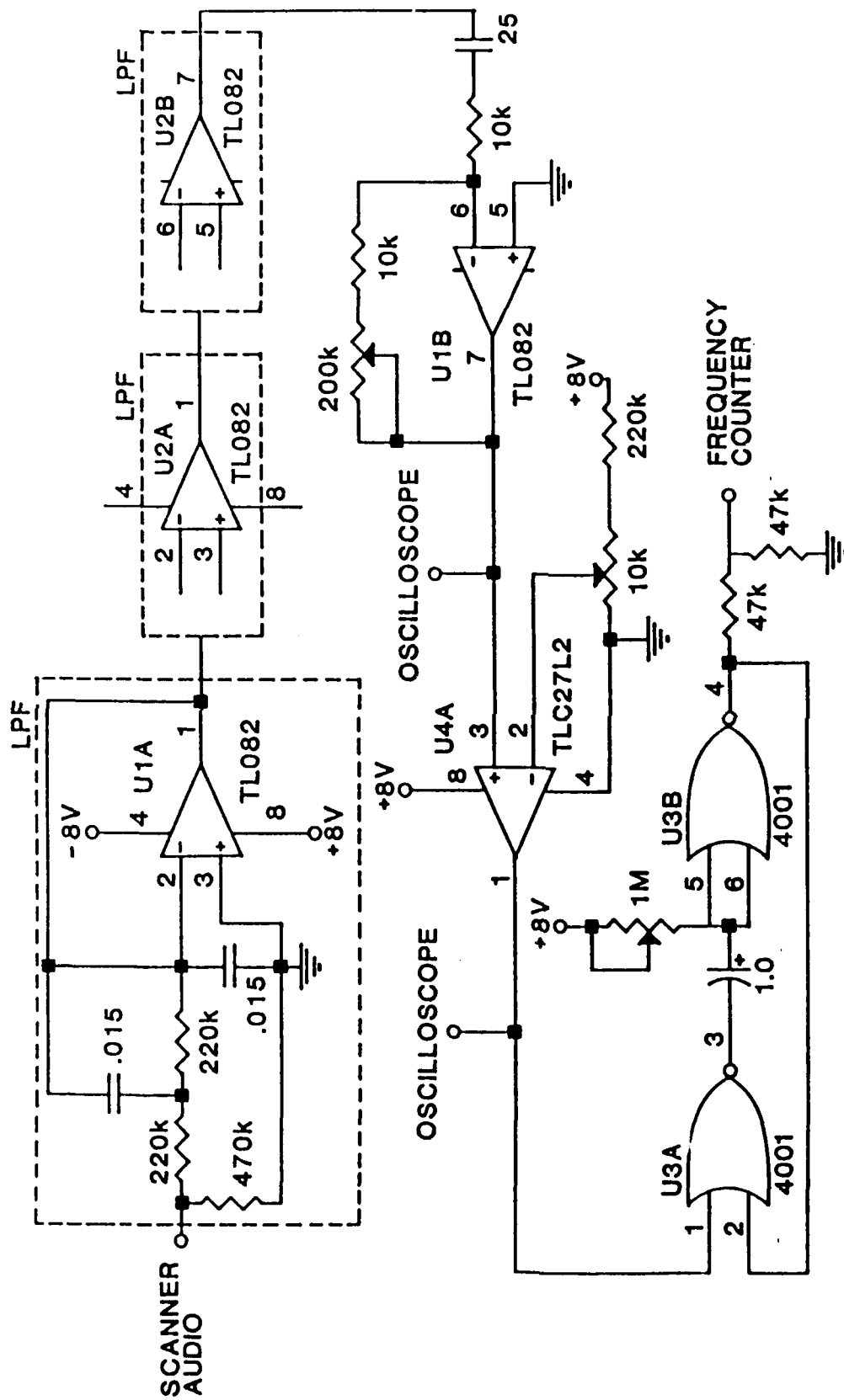


Figure III.3. LOW PASS FILTER AND ONE SHOT TRIGGER

carefully removed and the attachment sites cleaned with alcohol. Because the electrode attachment procedure can cause the skin to become irritated, a period of at least 3 days elapsed between heart rate measurements.

1. Heart Rate Monitor: Leads from the electrocardiocorder[®] were attached to silver/silver chloride electrocardiogram (ECG) monitoring electrodes (WA580 Foam Electrodes, MedTek, St Petersburg, FL) placed over the scapular and heart apex areas. On the left side, the first electrode was placed over the apex of the heart; the second over the caudal border of the scapula; and the third mid-way between the apex and scapula. On the right side, the first electrode was placed over the apex and the second over the caudal border of the scapula. The electrocardiocorder[®] was connected to the five leads and tested for operation using an electrocardiograph (Model EK/5A, Burdick Corp., Milton, WI) for stripchart ECG recordings. The clock inside the heart rate recorder was synchronized with the clock on the computer (IBM Model P/S 2-50) which accessed the radio signals from the radio transmitters.
2. Radio transmitter: Radio transmitter leads were kept clipped together to protect the amplifier section from overload by external fields. With the receiver/scanner tuned to the appropriate radio frequency, the transmitter was switched to the "on" position with the electromagnetic pulse generator (electromagnetic pulse generator) and the receiver/scanner was fine-tuned until no static (i.e., "white noise") was heard. The leads then were unclipped and connected to the ECG monitoring electrodes: the right lead over the right apex of the heart and the left lead over the caudal border of the scapula. Because the transmitter is "turned-off" when the leads are unclipped, it must be "turned-on" with the pulser. A storage oscilloscope (Model 7623A, Tektronix, Beaverton, OR) was used to observe ECG waveforms. After all transmitters were fitted and checked, the telemetry system was switched to sequentially monitored heart rates. Heart rate for each animal was counted for 10sec within a 20sec interval.

Data collection: After the radio transmitters and heart rate monitors were externally mounted, the animals were placed into their respective pens. Although heart rate data were collected for a minimum of 2 hr, analysis involved only a 30 min segment which began 60 min after the mounting of the radio transmitters and monitors. The 60 min

delay provided a sufficient period of time for the animals to adjust physiologically to the instrumentation and to obtain stable heart rate values.

Statistical analysis: Heart rate data were analyzed using a commercially available personal computer software program (Stat View 512; Winer, 1971).

C. RESULTS AND DISCUSSION

After placement of the heart rate measuring devices, the animals were returned to their respective pens. Although heart rates were recorded for a period of 2 hr, initial analyses involved both 30 min and 60 min mean heart rates taken from each animal following a 60 min period of readjustment to their pens and heart rate measuring devices. The relationship between the 30 min and 60 min mean heart rates obtained using the heart rate monitors is described in Figure III.4. Correlations also were made between the 30 min and 60 min mean heart rates obtained using radio transmitters (Figure III.5). These relationships suggest that either the 30 min or 60 min mean heart rate values could be used in subsequent analyses.

The relationship between 30 min mean heart rate values obtained from radio transmitters and monitors in control goats is described by the equation:

$$y = 0.808x + 23.934 \quad (1)$$

where:
 y = 30 min mean heart rate from radio transmitter, and
 x = 30 min mean heart rate from monitor

Further, the radio transmitter data were highly correlated ($r = 0.904$; $r^2 = 0.817$; $p < 0.0001$) with those obtained from the heart rate recorders. Although there is no consistent trend, the percent difference between these two devices (Table III.1) may be partly related to the fact that the radio transmitter signals represent time-averaged heart rates (i.e., over a 30-sec period) whereas the heart rate monitor records beat-to-beat heart rates. It also may be that with minor modifications of the transmitter circuitry, the percent differences could be reduced. A non-significant ($p > 0.05$) degradation of signal strength, proportional to operational use of the radio transmitter, was observed.

Figure III.4. The relationship between 30 min and 60 min mean heart rates obtained from heart rate monitors.

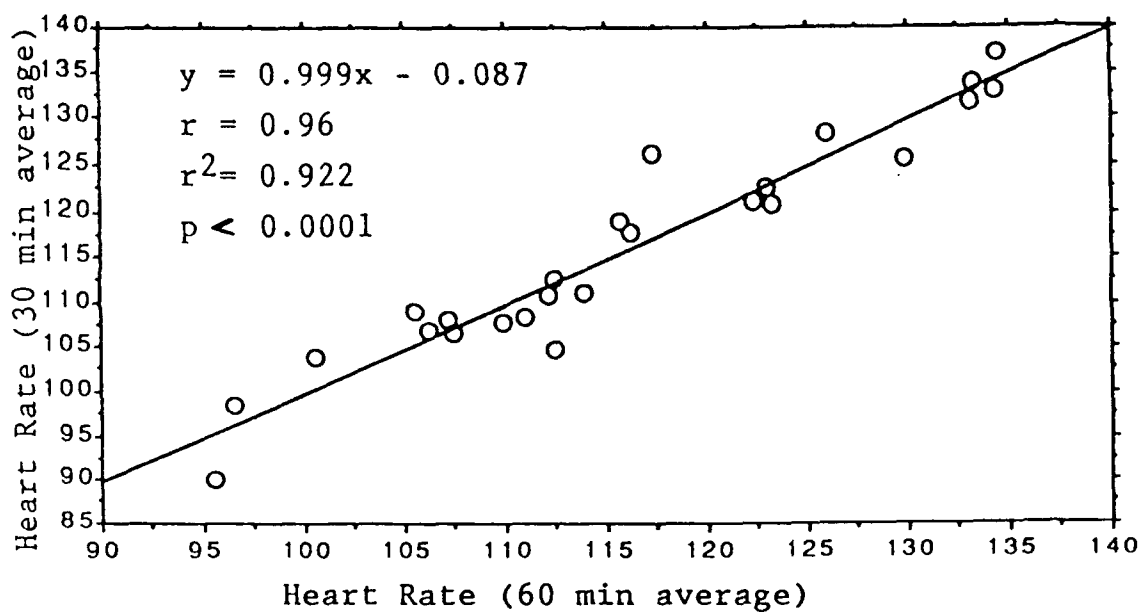


Figure III.5. The relationship between 30 min and 60 min mean heart rates obtained from radio telemetry.

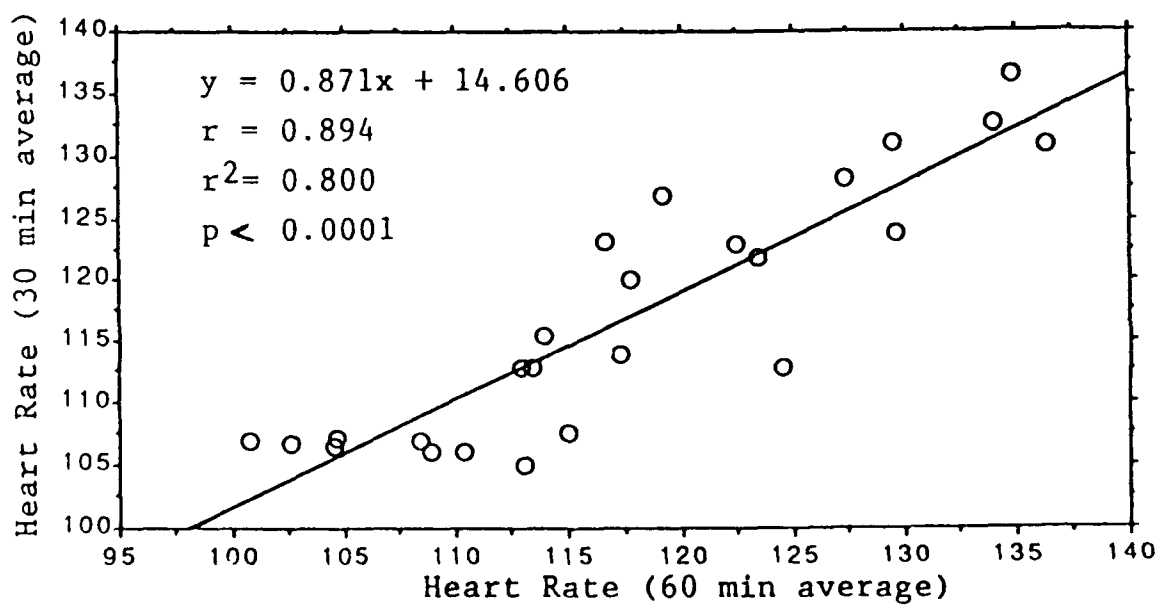


Table III.1. Comparison of heart rates (Mean \pm SE) obtained from four goats using two heart rate monitors and six radio transmitters for each animal.

Animal	n ¹	Radio transmitters	Heart Rate Monitors	Percent Difference ²
Goat No. 53	5	126.2 \pm 2.2	126.9 \pm 3.1	+0.6
Goat No. 57	8	116.7 \pm 2.5	112.9 \pm 3.4	-3.4
Goat No. 59	6	107.3 \pm 1.2	104.0 \pm 3.2	-3.2
Goat No. 60	6	119.6 \pm 6.2	122.4 \pm 5.2	+2.3
Mean of above groups		117.0 \pm 2.0	111.8 \pm 5.0	-4.6

¹ Number of observations

² $\frac{\text{Heart Rate Monitor Value} - \text{Transmitter Value}}{\text{Heart Rate Monitor Value}} \times 100$

Utilization of radio transmitters rather than heart rate monitors to measure heart rates of homoiothermic animals provides many advantages. For example, in comparing costs of the two techniques, transmitters can be fabricated in the laboratory at a unit cost of about \$25 whereas the heart rate monitors have a unit price of about \$1500. There is a similar cost differential for necessary support equipment. Thus, using radio transmitters, heart rates from a large number of animals can be obtained virtually simultaneously at a relatively low unit or total cost.

D. CONCLUSIONS

The usefulness of the locally designed and fabricated radio transmitter as a non-invasive technique to evaluate heart rates of domestic goats was successfully demonstrated through comparison with data recorded from a commercially available heart rate monitor. The descriptions of the telemetry system, radio transmitter, lowpass filter, and one-shot trigger also have been provided.

IV. OBJECTIVE NO. 2: HEART RATE AS PREDICTOR OF PLASMA CORTISOL LEVELS

A. INTRODUCTION

When animals are exposed to conditions for which they are neither accustomed nor physiologically adapted, they experience a stressor response which is characterized by elevated serum or plasma glucocorticoids. The elevation of glucocorticoids is a protective response of the body for mobilizing glucose to meet the physiological demands

induced by a stressor. Thus, plasma corticosteroids have been used as indicators of stressor response (Selye, 1950).

But, their reliability as stress indicators has been challenged because many factors can cause an elevation of these substances. For example, procedures associated with collecting blood samples (e.g., venipuncture, animal handling) have been shown to cause elevated level of plasma corticosteroids (Bassett and Hinks, 1969). Lymphopenia and neutrophilia also have been reported to be stressor responses (Dalton and Selye, 1939) but these values also have been reported to be influenced by administration of cortisol in the horse (Burguez *et al.*, 1983).

So, the purpose of the study reported here was to determine whether heart rate could be used as a predictor of stressor response in goats. Availability of such a predictor would enhance collection of data on the physiological consequence of exposing animals to stressors. Comparisons were made between heart rates and plasma cortisol levels and white blood cell changes in goats exposed to minimal (i.e., <50 dB(A)) and those exposed to elevated noise (i.e., 105 dB(A)) levels. Plasma cortisol was measured because it is considered to be the predominant corticosteroid in ovine (Bassett and Hinks, 1969) and caprine (Lindner, 1964).

B. MATERIALS AND METHODS

Animal environment, feeding and watering schedules, heart rate measurements, external mounting of transmitters, and telemetry system, were as described for the Objective entitled VALIDATION OF A HEART RATE RADIO TRANSMITTER (See Section III.B and III.C above). A total of 12 mixed-breed goats (6 male and 6 female) was used.

Noise Exposure: The goats were exposed to single event and multiple event A-weighted noise levels (dBA) as described in the Noise Simulation System (NSS) User's Manual (Chavez *et al.*, 1990). The single event Equivalent Sound Level (LEQ) for the experimental group was 105.3 dBA for the aircraft flyover signal ID No.4 (F4D). The control group noise levels did not exceed 50 dBA. All noise exposures were separated by at least 6 days. Multiple event exposures consisted of 3 single event at 6 min intervals or 13 single events at 1 min intervals over a total time period of 12 min. Following completion of the above, the location of the groups of animals was reversed and, after an interval of 10 da for conditioning the animals to their new pens (McNatty and Young,

1973), the sequence was replicated. Because of this crossover, there was a total of 12 control and 12 experimental animals.

Blood Sampling: Blood samples (~4cc) from the jugular vein were collected in vacutainers (Becton Dickinson, Reorder No 6450) containing EDTA as an anticoagulant. Samples were collected at about the same time intervals each day of collection and prior to exposing the goats to a noise stressor (Control Samples) and at ~20 min, ~75 min, and ~120 min following the noise stimulus (Experimental Samples). Immediately following collection, the vacutainers were placed in crushed ice until hematologic measurements were completed, usually within 45-min of collection.

Hematology: Packed cell volume (PCV) was determined using the microcapillary hematocrit tube method (Clay Adams, Autocrit Centrifuge, Model No. 0551) and plasma protein was measured using a hand-held refractometer (Schuco No. 21300). Blood smears were made by the pull slide technique and stained with anthene, thiazine, azure A, and methylene blue dyes (Harleco's Diff Quik Stain). Blood samples then were placed in a clinical centrifuge (Model CL, International Equipment Company) and the plasma separated from the formed elements. Plasma was transferred to 2 ml capped polypropylene vials (Naige Company, Catalog No. 5000-0020) which were placed in a freezer (-20°C) until plasma cortisol measurements were completed.

Radioimmunoassay: Radioimmunoassay of plasma cortisol was completed using a commercially available kit (Baxter Travenol Diagnostics, Inc., Cortisol Radioimmunoassay Kit).

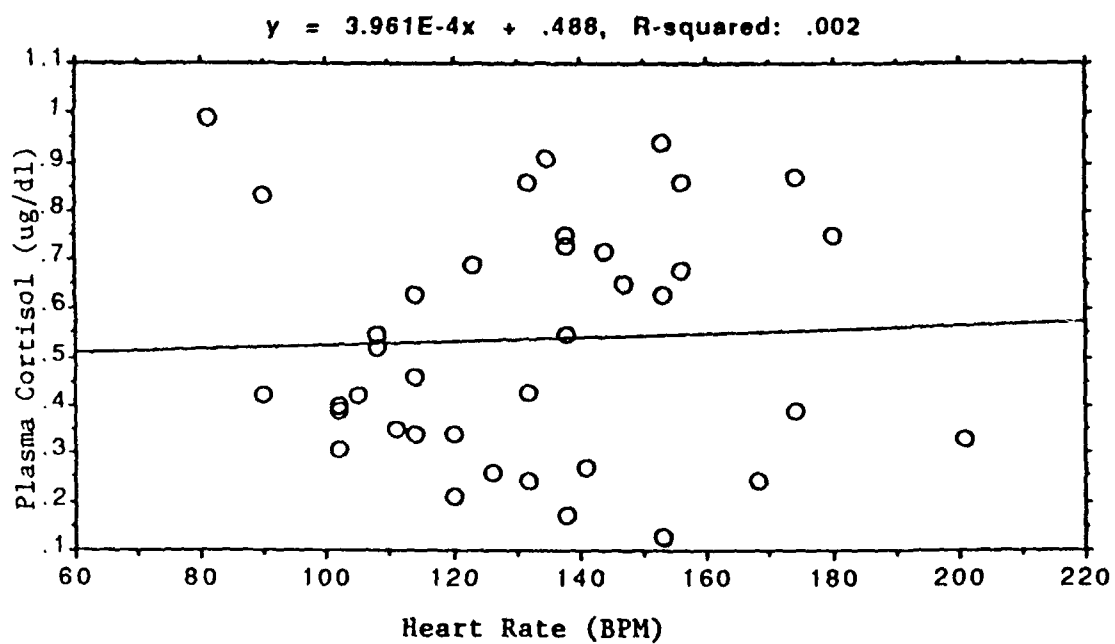
Statistical analysis: Data were analyzed using a commercially available personal computer software program (Stat View 512; Winer, 1971).

C. RESULTS AND DISCUSSION

Goat plasma cortisol was observed to be 0.90 ± 0.13 ug/dl. This value is qualitatively similar but less than 1.2 ug/100ml reported by Lindner *et al* (1964) and is about 10% of the plasma cortisol found in humans (Cohn and Kaplan, 1966).

A simple regression analysis revealed no statistically significant correlation between plasma cortisol and heart rates obtained from radio transmitters (Figure IV.1).

Figure IV.1. The relationship between plasma cortisol and heart rate



This finding conflicts with a previous report in which a linear relationship between heart rate and plasma cortisol was suggested for bighorn sheep (Harlow *et al.* 1987). However, in the latter study, the relationship was observed in individual vice group mean values as reported herein. Further, in the Harlow *et al.* (1987) study, blood samples were obtained via indwelling jugular vein cannulae. Because cannulae were used, blood samples could be obtained rapidly and at more frequent intervals (i.e., 5, 19, 15, 20, 30, 48, and 100 min intervals compared to only three intervals (i.e., 20, 75, and 120 min) reported herein.

A comparison of heart rate and plasma cortisol changes preceding and following exposure of goats to jet aircraft noise is displayed in Table IV.1. There were no detectable increases in either heart rate or cortisol for any of the post-noise exposure periods used except for the 75 min post-noise heart rate which is thought to be unrelated to the noise stimulus. Either the noise stimuli used in this study were not of sufficient intensity to elicit an adrenal response or the response was of such short duration that it was undetectable within the 20 min post-noise plasma sample.

It also may be that the high variability in plasma cortisol values obscured the possible relationship between heart rate and cortisol. It is not known whether that variability was due to the inherent characteristics of cortisol, radioimmunoassay technique relative to goat plasma, animal species, or some combination of those factors.

Table IV.1. Comparison of heart rate and plasma cortisol changes preceding and following exposure of goats to jet aircraft noise. All values are Mean \pm SE.

<u>Condition</u>	<u>n¹</u>	<u>Heart Rate</u>	<u>Cortisol</u>
Pre-noise	12	125 \pm 6	0.71 \pm 0.09
20 min Post-noise	14	137 \pm 8	0.61 \pm 0.12
75 min Post-noise	11	146 \pm 9 ²	0.70 \pm 0.16
120 min Post-noise	5	119 \pm 14	0.38 \pm 0.18

¹ Number of determinations from 9 goats

² p < 0.05 compared to pre-noise condition

D. CONCLUSIONS

Because of the advantages of using cardiac frequency as an indicator of plasma cortisol, further studies should be conducted to fulfill the intent of this objective. If additional studies are conducted, the protocols should provide for repetitive blood sampling from indwelling cannulae similar to the procedure employed by Harlow *et al* (1987).

V. OBJECTIVE NO. 3: BEHAVIORAL RESPONSES OF GOATS TO SIMULATED JET AIRCRAFT NOISE

A INTRODUCTION

There have been many reports on the behavioral responses of various domestic animals (Bond, 1971; Ewbank, 1977) and wildlife (Ames, 1978) to noise and sonic booms. In domestic animals, the representative behavioral signs include startle reflex, avoidance behavior, and vocal repetition rates (Espmark *et al*, 1974; Ruth, 1976). All species studied appear to readily adapt to sonic booms and display behavioral startle but no loss of production: cattle exhibit brief alterations in behavior, sheep temporarily stop grazing, ponies react to a greater extent than cattle, pigs show little response, and poultry display the most pronounced response of all observed domestic animals (Ewbank, 1977).

Heart rate changes and avoidance behavior have been used to evaluate goat encounters with people (Lyons and Price, 1987). Although heart rate telemetry is more objective, it also is more costly and behavioral cues have been found to be useful in assessing an animal's response to human disturbances. No responses to disturbances were detected using heart rate telemetry that were not apparent from behavioral cues alone (MacArthur *et al*, 1982).

The purposes of the study reported here were to describe the behavioral responses of domestic goats to single and multiple jet aircraft noise exposures and to determine whether the behavioral responses could be detected in the absence of measurable physiological responses.

B. MATERIALS AND METHODS

Animal environments, feeding and watering schedules, heart rate measurements, external mounting of transmitters, and telemetry system, all were as described for Objective No. 1 entitled VALIDATION OF A HEART RATE RADIO TRANSMITTER (See Section III.B above). Animals and noise exposures were as described for Objective No. 2 entitled HEART RATE AS A PREDICTOR OF PLASMA CORTISOL (See Section IV.B above).

Behavioral Responses: Behavior was recorded on black and white videotapes (Radio Shack, Supertape Videocassette, T-120) on each day the animals were exposed to a noise stimulus. The recording period began about 1 hr prior to and continued for at least 2.5 hr following the noise stimulus. Behavioral responses included three defined groups (a) no reaction, (b) moderate reaction, and (c) strong reaction (Table V.1).

Table V.1. Behavioral responses of goats (Espmark *et al.*, 1974).

Eating

No reaction	
Moderate reaction	1. Startle, no interruption of eating 2. Interruption of eating, raising of head, pricking of ears, no locomotion
Strong reaction	Interruption of browsing, locomotion of up to 10m

Lying

No reaction	
Moderate reaction	1. Intentions to get up. 2. Quick raising of head, pricking of ears
Strong reaction	Momentary rising into standing position

Standing

No reaction	
Moderate reaction	Startle, transient raising of head, no locomotion
Strong reaction	Startle, transient nodding of head, locomotion

Walking

No reaction	
Moderate reaction	Interruption of walking, raising of head, looking around
Strong reaction	Conversion of walking into running

Video Cameras: In each animal pen two video cameras (Koyo, TVC 4000 series, Closed Circuit TV Camera) were suspended from the ceiling so that each camera covered one-half of the pen area. Two clocks (Seth Thomas, Model No. 0709) also were suspended in such a manner that each camera was able to record the time of day that the behavior of the animals was videotaped. The month/day were printed in 5 in high block numerals on 14 cm X 21.5 cm sheets of paper which were attached to each clock.

Videocassette Recorders: The video cameras were wired to TV monitors (Magnavox, Model BX 3748) and video cassette recorders (Quasar, Model VH 5290). Thus, behavior could be monitored in 'real time' and videotaped for viewing at a later time.

C. RESULTS AND DISCUSSION

Analysis of video tape recordings of goat behavior revealed that goats from both experimental groups displayed behavioral responses to single and multiple event noise exposures (Table V.2). No behavioral responses to single or multiple event noise exposure were observed in either control group of animals.

Table V.2. Comparison of responses to single event and multiple event noise stimuli. Thirteen goats were included in each experiment group except for multiple 3 in Experiment No. 2 in which only 12 goats were used.

Experiment Group	Type	# of Events ¹	Interval (da) ²	Noise Stimuli Event	
				Response to Stimuli Number	
				First	Subsequent ³
No. 1	Single	1	0	Strong	---
	Single	1	8	Strong	---
	Multiple	3	8	Strong	None
	Multiple	13	6	Strong	Moderate
	Single	1	10	Moderate	---
No. 2	Multiple	13	0	Strong	Moderate
	Single	1	10	Strong	---
	Multiple	3	14	Strong	Strong

¹ Number of events in a 12min multiple type noise exposure

² Interval between noise stimuli event; 0 interval means the first noise exposure for "noise-naive" animals

³ Refers to all stimuli following the initial stimuli in a multiple event exposure

When animals were exposed to a single event noise stimulus, the behavioral response was listed as "strong" (Table V.1) for all animals in both experiment groups except for the "moderate" response in the third single event exposed goats in Experiment 1. It also should be noted that the Multiple 3 animals in Experiment No. 1 displayed no response to the second and third noise stimuli yet in the Multiple 3 goats of Experiment 2 there was a "strong" response to the second and third stimuli. These differences in response are thought to be related to habituation which involves, among other things, both the interval between noise exposures and number of repetitive exposures to the noise stimuli.

An analysis of heart rate data (Table V.3) confirms the observed behavioral changes. Although heart rates increased in both control and experimental groups, the increases were 55-93% greater in the latter group.

Table V.3. Heart rate changes (Mean \pm SE) in control and experimental goats in the 5 min periods immediately prior to and following application of the single and multiple event noise stimuli. Heart rates were recorded with a heart rate monitor (See text).

Stimulus	n ¹	Control		% Incr ²	Experimental		% Incr ²
		Pre-noise	Post-noise		Pre-Noise	Post-noise	
Single Event	3	107 \pm 9	114 \pm 7	6.5	129 \pm 30	142 \pm 33	10.1
Multiple Event	4	124 \pm 5	126 \pm 8	1.6	129 \pm 13	139 \pm 13	7.8

¹ Number of animals

² $\frac{\text{Post-noise} - \text{Pre-noise}}{\text{Pre-noise}} \times 100$

Because the single and multiple event noise exposures occurred at different time intervals (Table V.4), an effort also was made to determine the relationship between the length of the interval and the net percent increase in heart rate in pre- and post-noise exposed control and experimental goats. A regression analysis revealed a non-significant indirect correlation ($r = 0.333$; $r^2 = 0.111$; $p < 0.52$) between the length of noise exposure interval and the net percent increase in heart rate.

Table V.4. Relationship between the net percent increase in heart rate with the time interval between noise exposures. The number of events in a 12 min noise exposure and the interval descriptions are the same as for Figure V.3 above.

Date	Type Event	Interval (da)	Net % Increase
3 July	Single	0	6.5
11 July	Multiple	8	1.2
17 July	Multiple	6	0.0
27 July	Single	10	3.3
3 August	Multiple	0	7.3
13 August	Single	10	--
27 August	Multiple	14	3.3

D. CONCLUSIONS

The noise stimuli used in this study resulted in measurable behavioral responses in goats. The degree of response appears to be related to habituation--both length of

interval between responses and the number of repetitive exposures. Further, the heart rate data revealed no responses that were not apparent from changes in behavior.

VI. OBJECTIVE NO. 4: PHYSIOLOGICAL RESPONSES OF GOATS TO SIMULATED JET AIRCRAFT NOISE

A. INTRODUCTION

Changes in adrenocortical activity (Selye, 1950), body temperature (Stephens, 1980; Stermer *et al.*, 1980), respiratory frequency (Stephens, 1980), and heart rate (MacArthur *et al.*, 1982) all have been used to measure the response of domestic animals to environmental stressors. Because animal handling--which often is associated with sample collection--will cause an increase in all of these measurements, remote monitoring of vital signs has been suggested as a more reliable method of evaluating the physiological status of an animal (Stermer *et al.*, 1980).

Behavioral adjustments to environmental stressors also have been reported to provide information on the physiological well-being of animals (Dantzer and Mormede, 1983). But, behavioral changes do not always give a reliable indication of potential physiological changes. For example, Lyons and Price (1987), using dam-reared and human-reared goats observed no consistent relationship between heart rate and behavior. Moreover, they reported that significant behavioral changes were not accompanied by group differences in heart rate.

So, an important question is "What parameters can be used to evaluate the effects of environmental stressors?" Utilizing both radiotelemetry to monitor body function (e.g., heart rate and body temperature) and other standard blood tests (e.g., plasma cortisol, relative numbers of lymphocytes and neutrophils), the responses of goats to simulated jet aircraft noise were evaluated and the results are reported here.

B. MATERIALS AND METHODS

Animal environments, feeding and watering schedules, heart rate measurements, external mounting of transmitters, and telemetry system, all were as described for Objective No. 1 entitled VALIDATION OF A HEART RATE RADIO TRANSMITTER (See Section

III.B above). Animals and noise exposures were as described for Objective No. 2 entitled HEART RATE AS A PREDICTOR OF PLASMA CORTISOL (See Section IV.B above).

Statistical analysis: White blood cell, heart rate, and plasma cortisol changes were analyzed using a commercially available personal computer software program (Stat View 512; Winer, 1971).

C. RESULTS AND DISCUSSION

There were no detectable changes in plasma cortisol for any of the goat groups or noise stressors (Table VI.1). This finding is similar to that reported by Besch (1989) but unlike the one reported by Ames (1972) for lambs. The lack of response in goats exposed to the jet aircraft engine noise stressor may have been partly due to the fact that the intensity/duration of the stimulus was insufficient to elicit a response or that the response was of such short duration (i.e., it was transient) that a response was undetectable within the 20 min post-noise plasma sample. It also is possible that the goat does not respond in any detectable way--other than behavioral--to a noise stimulus from jet aircraft engine noise. If so, this indeed would be a unique animal model for noise studies.

Table VI.1. Plasma cortisol values (ug/dl) in goats prior to and following exposure to single and multiple noise stimuli (See text for details). All values are Mean \pm SE.

Group	n	Stimulus	Pre noise	Post-noise	
				~20 min	~75 min
Control					
	14	Single	0.83 ± 0.09	0.85 ± 0.08	0.97 ± 0.12
	7	Multiple-3 ¹	1.28 ± 0.34	0.88 ± 0.33	0.56 ± 0.25
	7	Multiple-13 ²	0.78 ± 0.25	0.42 ± 0.08	0.68 ± 0.15
Experimental					
	12	Single	0.78 ± 0.10	0.82 ± 0.09	0.68 ± 0.10
	6	Multiple-3 ¹	1.46 ± 0.40	1.00 ± 0.52	0.55 ± 0.14
	6	Multiple-13 ²	1.39 ± 0.62	1.01 ± 0.39	0.67 ± 0.18

¹ Three single event stimuli in 12 min; ² 13 single event stimuli in 12 min

Although no consistent, significant changes in relative numbers of lymphocytes (Table VI.2) or neutrophils (Table VI.3) were detected, there was a tendency toward lymphopenia and neutrophilia in the experimental group of goats.

Table VI.2. Lymphocyte (%) changes (Mean \pm SE) in goats exposed to single and multiple event noise exposures (See text).

	n ¹	Pre-noise	n ¹	Post-noise		
				+20 min	n ¹	+75 min
Control:						
Single Event	18	60.2 ± 2.1	18	60.5 ± 2.2	18	60.1 ± 2.5
Multiple Event ²	11	64.4 ± 1.9	12	62.5 ± 2.5	13	61.1 ± 3.3
Multiple Event ³	13	58.7 ± 3.0	12	60.5 ± 2.3	13	57.8 ± 3.4
Jain (1986)		56.0				
Experimental:						
Single Event	18	56.7 ±2.3	18	60.4 ± 2.0	18	57.9 ± 2.4
Multiple Event ²	12	58.8 ± 3.7	12	57.9 ± 4.8	12	56.7 ± 4.2
Multiple Event ³	13	56.4 ± 3.2	13	54.2 ± 3.4	13	57.9 ± 2.7

¹ Number of animals

² Three single event stimuli in 12 min

³ Thirteen single event stimuli in 12 min

Table VI.3. Neutrophil (%) changes (Mean \pm SE) in goats exposed to single and multiple event noise exposures (See text).

	n ¹	Pre-noise	Post-noise			
			n ¹	+ 20 min	n ¹	+75 min
Control:						
Single Event	17	35.8 ± 2.0	17	34.0 ± 2.3	17	37.1 ± 2.7
Multiple Event ²	12	32.5 ± 3.9	12	36.8 ± 2.3	12	38.5 ± 3.4
Multiple Event ³	12	33.2 ± 3.3	12	31.4 ± 3.1	12	36.8 ± 3.2
Jain (1986)		36.0				
Experimental:						
Single Event	18	36.6 ± 3.6	17	37.2 ± 2.0	17	40.9 ± 2.3
Multiple Event ²	12	36.2 ± 3.7	12	40.0 ± 2.9	12	39.4 ± 2.7
Multiple Event ³	12	37.0 ± 3.1	12	38.2 ± 4.5	12	38.9 ± 3.8

¹ Number of animals

² Three single event stimuli in 12 min

³ Thirteen single event stimuli in 12 min

Lymphopenia and neutrophilia have been reported to be stressor responses (Dalton and Selye, 1939) and the findings reported here suggest that the noise stimuli were not of sufficient intensity/duration to cause a statistically significant response. Whether these findings can be extrapolated to other species of domestic animals is uncertain. Nonetheless, there is evidence to assume that occasional flyovers of jet aircraft should not pose a particularly "stressful" environment for domestic goats.

No statistically significant differences were detected in packed cell volumes, plasma protein, or growth rate between control and noised exposed groups.

D. CONCLUSIONS

Because no hematological or humoral responses were detected, neither the noise intensities nor durations used in this study appear to be physiologically "stressful" to goats. Further, if the physiological responses are of such short duration that they are undetectable 20 min after exposure to a noise stressor, then behavioral changes, if any, would appear to be good indicators of an animal's response to acute stress.

VII. RECOMMENDATIONS

A. The lack of detectable responses to jet aircraft noise in goats appears to be related to the intensity or duration of the noise stressor. Further studies are indicated to determine the noise threshold that will elicit detectable/measurable responses of goats.

B. If additional studies are conducted, radiotelemetry should be used to obtain heart rates and indwelling venous cannulae should be used in the collection blood samples from domestic animals exposed to noise stressors.

C. Because of the advantages of using cardiac frequency as an indicator of plasma cortisol during exposure to environmental stressors, additional research should be conducted to determine if this relationship can be established for goats.

D. The behavioral responses of goats to simulated jet aircraft engine noise suggests that either habituation or repeated exposures modifies the animals' responses to subsequent noise stressors. What is unknown is the temporal relationship between

exposures to a noise stressor(s) and elicitation of a detectable response. In any event, additional study in this area also should be encouraged.

E. Data collected during this study should be published in the open scientific literature. In the meantime, a copy of this report should be shared with the Noise and Sonic Boom Impact Technology (NSBIT) program office at Wright-Patterson AFB, Ohio.

F. Because animal handling can affect the measured physiological parameters, studies using non-invasive procedures should be encouraged.

VIII. REFERENCES

1. Ames, D. R. Physiological Responses to Auditory Stimuli. In: Effects of Noise on Wildlife. J. L. Fletcher and R. G. Busnel, Eds., New York: Academic Press, 1978, pp 23-45.
2. Ames, D. R. and L. A. Arehart. Physiological response of lambs to auditory stimuli. J Anim Sci 34(6): 964-998, 1972.
3. Anonymous. Guide for the Care and Use of Agricultural Animals in Agricultural Research and Teaching. 1st Ed. Consortium for Developing a Guide for the Care and Use of Agricultural Animals in Agricultural Research and Teaching. Washington, DC: National Association of State Universities and Land-Grant Colleges, 1988, 74p.
4. Bassett, J. M. and N. T. Hinks. Micro-determination of corticosteroids in ovine peripheral plasma: effects of venipuncture, corticotrophin, insulin and glucose. J Endocrinol, 1969, Vol 44(3), pp 378-403.
5. Bell, W. B. Animal responses to sonic booms. J Acoust Soc Amer, 1972, Vol 51, pp 758-765.
6. Besch, E. L. FINAL REPORT to Universal Energy Systems, Incorporated, Dayton, OH, 1989 USAF-UES Summer Faculty Research Program (SFRP), Tyndall AFB, FL, entitled "Effect of Jet Aircraft Noise on Domestic Goats", 8 Sep 1989.

7. Bond, J. Noise--Its effect on the physiology and behavior of animals. Agric Sci Rev, 1971, Vol 9(4), pp 1-10.
8. Burguez, P. N., J. Ousey, R. S. G. Cash, and P. D. Rosedale. Changes in blood neutrophil and lymphocyte counts following administration of cortisol to horses and foals. Equine Vet J, 1983, Vol 5(1), pp 58-60.
9. Casady, R. B. and R. P. Lehmann. Response of farm animals to sonic booms. Studies at Edwards Air Force Base, California, 6-30 June 1966. Interim Report. Beltsville, MD: US Department of Agriculture, 1967, 8 p.
10. Chavez, P., S. Tomooka, J. A. Ciarletta, and T. Howe. Noise simulation system for low-level aircraft overflights: Effects on pregnant mares. A User's Manual. BBN Report No. 7243, Job No. 609001. BBN Systems and Technologies Corporation, 1990.
11. Cohn, C. and A. Kaplan. Blood Chemistry. In: A Textbook of Clinical Pathology. 7th Ed., S. E. Miller, ed. Baltimore: The Williams and Wilkins Company, 1966, pp 257-335.
12. Cottereau, P. Effect of sonic boom from aircraft on wildlife and animal husbandry. In: Effects of Noise on Wildlife. J. L. Fletcher and R. G. Busnel, eds. New York: Academic Press, 1978, pp 63-79.
13. D. J. and H. Selye. The blood picture during the alarm reaction. Folia Haematol, 1939, Vol 62, pp 397.
14. Dantzer, R. and P. Mormede. Stress in farm animals: a need for reevaluation. J Anim Sci, 1983, Vol 57, pp 6-18.
15. Espmark, Y, L. Falt and B. Falt. Behavioral responses in cattle and sheep exposed to sonic booms and low-altitude subsonic flights. Vet Rec, 1974, Vol 94(6), pp 106-113.
16. Ewbank, R. The effects of sonic booms on farm animals. Vet Annual, 1977, Vol 17, pp 296-306.

17. Harlow, H. J., E. T. Thorne, E. S. Williams, E. L. Belden, and W. A. Gern. Cardiac frequency: a potential predictor of blood cortisol levels during acute and chronic stress exposure in Rocky Mountain bighorn sheep (Ovis canadensis canadensis). Can J Zool, 1987, Vol 65, pp 2028-2034.
18. ILAR (Institute of Laboratory Animal Resources). Committee on the Care and Use of Laboratory Animals, Guide for the Care and Use of Laboratory Animals. NIH Publ 85-23, Washington, DC; Department of Health and Human Services, 1985, 85p.
19. Jain, N. C. Schalm's Veterinary Hematology, 4th Ed., Chapter 8. Philadelphia: Lea and Febiger, 1986, pp 225-239.
20. Lindner, H. R. Comparative aspects of cortisol transport: lack of firm binding to plasma proteins in domestic ruminants. J. Endocrinol, 1964, Vol 28(3), pp 301-320.
21. Lyons, D. M. and E. O. Price. Relationship between heart rates and behavior in goats in encounters with people. Appl Anim Behav Sci 1987, Vol 18, pp 363-369.
22. MacArthur, R. A., V. Giest, and R. H. Johnston. Cardiac and behavioral response of mountain sheep to human disturbance. J Wildl Manage, 1982, Vol 46(2), pp 351-358.
23. McNatty, K. P. and A. Young. Diurnal changes of plasma cortisol levels in sheep adapting to a new environment. J Endocrinol, 1973, Vol 56, pp 329-330.
24. Richards, J. I. and P. R. Lawrence. The estimation of energy expenditure from heart rate measurement in working oxen and buffalo. J Agri Sci, 1984, Vol 102, pp 711-717.
25. Ruth, J. S. "Reaction of arctic wildlife to gas pipeline related noise." J Acoust Soc Amer, 1976, Vol 60 (Suppl 1), S67 (Abstract).
26. Selye, H. The Physiology and Pathology of Exposure to Stress. Montreal: ACTA, Inc., Medical Publishers, 1950, 822p.

27. Shotton, L. R. Response of wildlife and farm animals to low-level military jet overflight. Reporter, 1982, Vol II(6), pp 161-164.
28. Stephens, D. B. Stress and its measurement in domestic animals: a review of behavioral and physiological studies under field and laboratory situations. Advances in Vet Sci Comp Med, 1980, Vol 24, pp 179-210.
29. Stermer, R. A., T. H. Camp, and L. R. Smith. Telemetry systems for monitoring physiological responses of cattle. Trans ASAE, 1980, Vol 23(1), pp 144-149.
30. Varosi, S. M., R. L. Brigmon and E. L. Besch. A simple remote-controlled power switch for internalized bioelectronic instrumentation. IEEE Biomed Eng 36(8): 858-860, 1989.
31. Varosi, S. M., R. L. Brigmon and E. L. Besch. A simplified telemetry system for monitoring body temperature in small animals. Lab Anim Sci 40(3): 299-302, 1990.
32. Winer, B. J. Statistical Principles of Experimental Design. New York: McGraw-Hill, 1971, 907p.
33. Zern, J. D. FINAL REPORT to Universal Energy Systems, Incorporated, Dayton, OH, 1989 USAF-UES Graduate Student Research Program (GSRP), Tyndall AFB, FL, 14 Aug 89

Migration of Organic Liquid Contaminants Using Measured and Estimated Transport Properties

**Dr. Avery H. Demond, Assistant Professor
Department of Civil and Environmental Engineering
The University of Michigan
Ann Arbor, MI 48109-2125**

**Final Report
Research Initiation Program
Universal Energy Systems
4401 Dayton-Xenia Road
Dayton, OH 45432**

Sponsored by the Air Force Office of Scientific Research

June, 1991

Migration of Organic Liquid Contaminants Using
Measured and Estimated Transport Properties

by

Dr. Avery H. Demond, Assistant Professor

ABSTRACT

The Air Force has identified spills of jet fuel as a major source of groundwater contamination on bases across the nation. To investigate the separate phase movement of organic liquids such as jet fuel, numerical models are often used. Unfortunately, there are few data available for key relationships in these models for many organic liquids. Consequently, estimation techniques are often employed to determine the relationships of capillary pressure and relative permeability. Previous research has demonstrated that estimates of these relationships generated with common techniques sometimes do not correlate well with measurements. Thus, the use of these techniques may introduce error into simulations of jet fuel migration in the subsurface. The purpose of this research was to assess the magnitude of this error. Simulations of organic liquid flow were carried out using measured and estimated capillary pressures and relative permeabilities. A comparison of the simulations shows that the advance of the organic liquid front and the total mass of organic liquid infiltrated is significantly overpredicted when estimated properties are used. Thus, caution must be exercised in basing estimates of quantity of jet fuel spilled and location of the spill boundaries on simulations using estimates of the transport properties, capillary pressure and relative permeability.

The results of this research have been presented at the Symposium on Characterization of Transport Phenomena in the Vadose Zone, Tucson, AZ, April 2-5, 1991 (sponsored by American Geophysical Union and Soil Science Society of America), and will be submitted shortly for publication in Journal of Contaminant Hydrology.

ACKNOWLEDGEMENTS

I thank the Air Force Systems Command, Air Force Office of Scientific Research, Universal Energy Systems, and the Air Force Engineering and Services Center (Tyndall AFB) for their sponsorship and administration of this research. Others at the Air Force Engineering and Services Center should be mentioned for their part in obtaining sponsorship for this project: Mr. Jack Milligan, Dr. Daniel Stone, Dr. Tom Stauffer and Lt. Col. Tom Lubozynski. I also thank Jen-Mu Tang who was responsible for much of the computer work performed in the course of this research.

I. INTRODUCTION

The Air Force has identified spills of jet fuels, such as JP-4, as a major source of groundwater contamination on bases across the nation. Owing to the low aqueous solubility of jet fuel in water, jet fuel persists as a separate liquid phase and is transported as such in groundwater. Despite the recognition that multiphase flow processes are important, many questions about the physics of such processes remain unanswered.

To investigate the subsurface movement of organic liquids such as JP-4 from spill sites, often numerical models are used. These numerical models are based on equations which require the use of constitutive relationships for solution (Abriola and Pinder, 1985; Faust, 1985; Osborne and Sykes, 1986; Corapcioglu and Baehr, 1987; Kuppusamy et al., 1987). Unfortunately, little data exist for several of these constitutive relationships for the organic liquids of interest, making the accurate simulation of their migration difficult (Schwille, 1984; Faust, 1985; Pinder and Abriola, 1986). Two relationships of particular importance are capillary pressure and relative permeability as functions of saturation. Because of the lack of data, many numerical simulations employ relationships that have been estimated using a variety of techniques developed in soil science and petroleum engineering. However, as Demond and Roberts (1991a, 1991b) pointed out, the estimates produced by these techniques sometimes do not correlate well with measurements. Consequently, use of these techniques may introduce error into simulations of JP-4 migration in the subsurface.

II. OBJECTIVE

The purpose of this research was to evaluate the significance of the error introduced into simulations of two-phase flow through the use of estimated transport relationships rather than measured. To accomplish this purpose, two series of numerical simulations were carried out: one series using common estimation techniques, and the other using published capillary pressure and relative permeability data. Then, the two series of simulations were compared in terms of location of organic liquid displacement front and total mass of organic liquid infiltrated and the accuracy of the estimation techniques critiqued based on the comparison.

III. BACKGROUND

For a system composed of a rigid, non-deformable, homogeneous porous medium, and two mutually saturated, incompressible liquid phases, with no internal sources or sinks, the fundamental mass balance equations for two-phase flow may be written:

$$\nabla \cdot \left[\frac{\rho_w k k_{rw}}{\mu_w} (\nabla P_w - \rho_w g \nabla h) \right] = \frac{\partial(n \rho_w S_w)}{\partial t} \quad [1]$$

$$\nabla \cdot \left[\frac{\rho_n k k_{rn}}{\mu_n} (\nabla P_n - \rho_n g \nabla h) \right] = \frac{\partial(n \rho_n S_n)}{\partial t} \quad [2]$$

where

- ρ = density,
- k = intrinsic permeability,
- k_r = relative permeability,
- μ = viscosity,
- P = pressure,
- g = gravitational acceleration constant,
- h = elevation above a datum,
- n = porosity,
- S = saturation,
- t = time,

subscripts

- n = nonwetting (here, the organic liquid), and
- w = wetting (here, the aqueous phase).

To obtain a solution, the constitutive relationships of capillary pressure:

$$P_n - P_w = P_c = f(S) \quad [3]$$

where P_c = capillary pressure

and relative permeability:

$$k_{ri} = k_i/k = f(S) \quad [4]$$

where k_i = effective permeability to i th phase,

must be provided in order to form a close system of equations.

These relationships may be measured in the laboratory. But, because their measurement can be time-consuming or impractical, they are often estimated. Currently, the method of choice for estimating the primary drainage capillary pressure-saturation relationship for organic liquid-water systems follows from Leverett's function (Leverett, 1941) (Lenhard and Parker, 1987; Parker et al., 1987; Cary et al., 1989):

$$J(S_w) = P_{c1}/\gamma_{ORG1/H2O}(k_1/n_1)^{1/2} = P_{c2}/\gamma_{ORG2/H2O}(k_2/n_2)^{1/2} \quad [5]$$

where $J(S_w)$ = Leverett's function,
 $\gamma_{ORG/H2O}$ = interfacial tension between organic liquid and aqueous phases,

In situations where the systems have the same intrinsic permeability and porosity ($k_1=k_2$; $n_1=n_2$), then Leverett's function simplifies to:

$$J(S_w) = P_{c1}/\gamma_{ORG1/H2O} = P_{c2}/\gamma_{ORG2/H2O}$$

or, $P_{c2} = P_{c1} (\gamma_{ORG2/H2O}/\gamma_{ORG1/H2O}) \quad [6]$

Hence, the capillary pressure at a given saturation for an organic liquid-water system, can be generated from, for example, the air-water capillary pressure-saturation relationship for that soil by knowing the ratio of the liquid-liquid interfacial tensions. This method is usually applied to drainage, but it seems that it can be readily applied to imbibition if the relevant data are available (Demond and Roberts, 1991a). Eqn. [6] implies that the value for residual saturation of the wetting phase is the same for all systems; simply the capillary pressure at which that value is obtained varies.

Methods to estimate relative permeability can be divided into two basic groups: those based on capillary pressure measurements and those whose functional form is a power function of effective saturation, where effective saturation is defined as:

$$S_e = \frac{(S_w - S_{wr})}{(1 - S_{wr})} \quad [7]$$

where S_{wr} = residual saturation of the wetting phase

Techniques commonly used fall in both groups: Burdine's (1953) and Mualem's (1976), from the former, and Corey's (1954) and Wyllie's (1962) from the latter (Table 1). Both types of techniques have drawbacks. The power function models give the same results for all organic liquid-water pairs. Using the models based on capillary pressure measurements assumes that such data are available. If these data are estimated using Eqn [6], then again the same results are obtained for all organic liquid-water pairs. These methods were developed for drainage processes. They may be applied to the imbibition of the wetting phase in strongly-wetted systems, since that phase does not exhibit hysteresis. However, the nonwetting phase does display hysteresis. Yet, comparable methods for the prediction of k_{rn} during imbibition are uncommon.

Demond and Roberts (1991a, 1991b) recently compared estimated and measured capillary pressure and relative permeability relationships for organic liquid-water systems. Their study showed that as the interfacial forces decreased, the deviations between measured capillary pressure relationships and those estimated using Eqn. [6] increased (Fig. 1). The imbibition capillary pressure relationship showed a stronger effect than did the drainage relationship. The relative permeability of the wetting phase in strongly-wetted systems is reasonably well estimated by the models listed in Table 1 (Fig. 2). On the other hand, the relative permeability to the nonwetting phase is grossly overestimated particularly at low aqueous phase saturations (Fig. 3). In addition, the accuracy of the estimates for both phases grows poorer with decreasing interfacial forces.

Despite the observation of these differences between measured and estimated transport relationships, no quantitative discussion of the implications for the modeling of organic liquid migration was given. Notwithstanding the magnitude of these differences, they are only important in so far as they affect estimates of organic liquid movement in the subsurface. This research addresses that issue. To assess the effect of estimated transport relationships on predictions of two-phase flow, the migration of an organic liquid was simulated using both measured and estimated data. Three series of simulations were performed: using all measured data, using measured capillary pressure and estimated relative permeability relationships, and using estimates for both relationships.

TABLE 1. Correlations for Estimation of Drainage Relative Permeabilities

	$k_{rw} (-)$	$k_m (-)$
Burdine (1953)	$S_e^2 \frac{\int_0^{S_e} \frac{1}{P_c^2} dS_e}{\int_0^1 \frac{1}{P_c^2} dS_e}$	$\frac{\int_0^{S_e} \frac{1}{P_c^2} dS_e}{\int_0^1 \frac{1}{P_c^2} dS_e} (1 - S_e)^2 [(1 - S_e^{1/4})^4]$
Mualem (1976)	$S_e^{1/2} \left[\frac{\int_0^{S_e} \frac{1}{P_c} dS_e}{\int_0^1 \frac{1}{P_c} dS_e} \right]^2$	$\frac{\left[\frac{\int_0^{S_e} \frac{1}{P_c} dS_e}{\int_0^1 \frac{1}{P_c} dS_e} \right]^2}{(1 - S_e)^{1/2}} (1 - S_e)^{1/2} [(1 - S_e^{1/m})^m]^2$
Corey (1954)	S_e^4	$(1 - S_e)^2 (1 - S_e^2)$
Wyllie (1962)	S_e^3	$(1 - S_e)^3$

where $m = 1 - 1/n$ and $q = 1 - 2/n$ where n is determined by fitting Eqn. [10] to the drainage capillary pressure-saturation relationship.

Figure 1. Comparison of measured and estimated capillary pressure-saturation relationships.

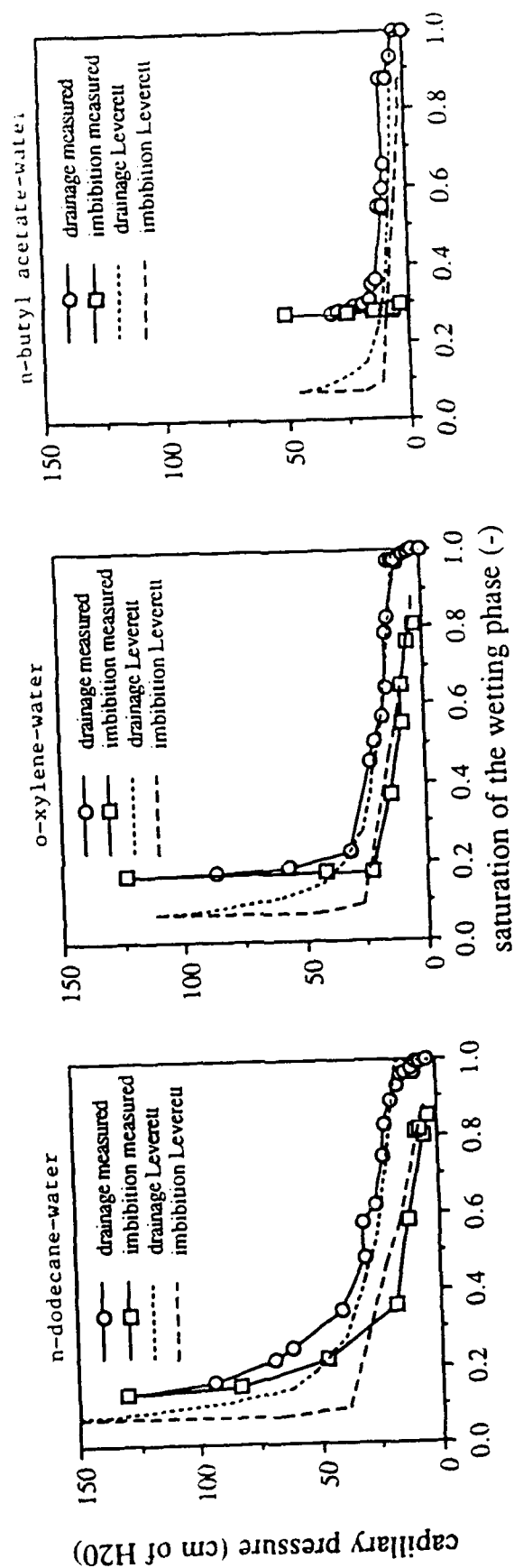


Figure 2. Comparison of measured and estimated relative permeability-saturation relationships for the aqueous phase.

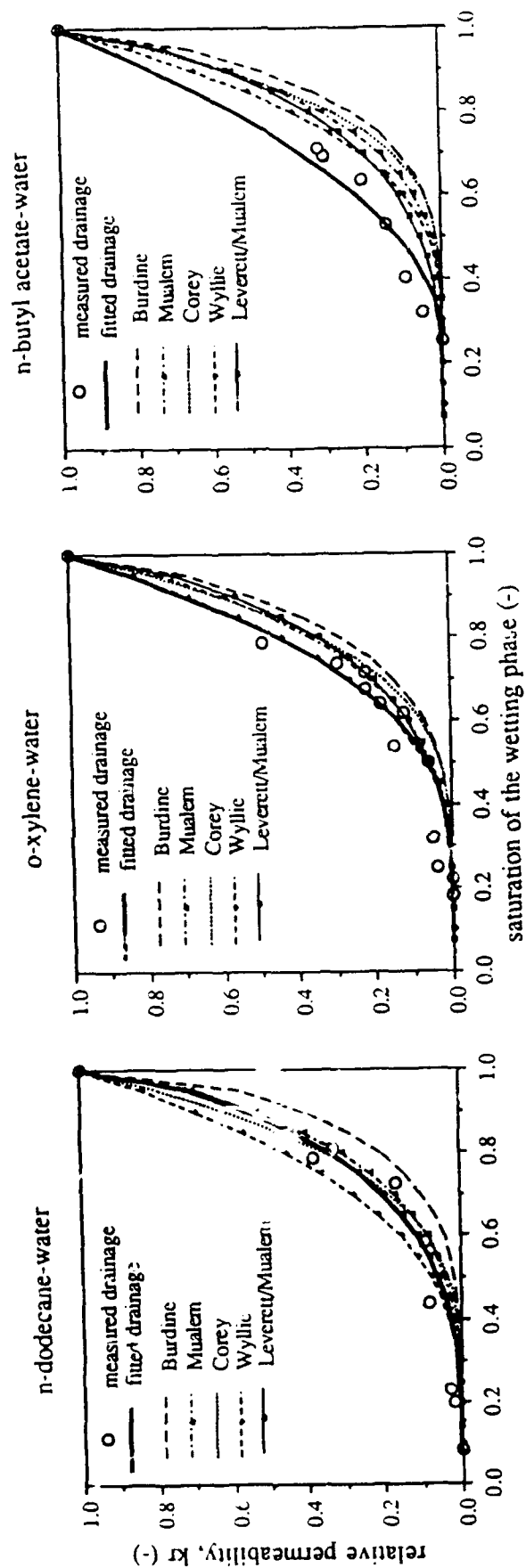
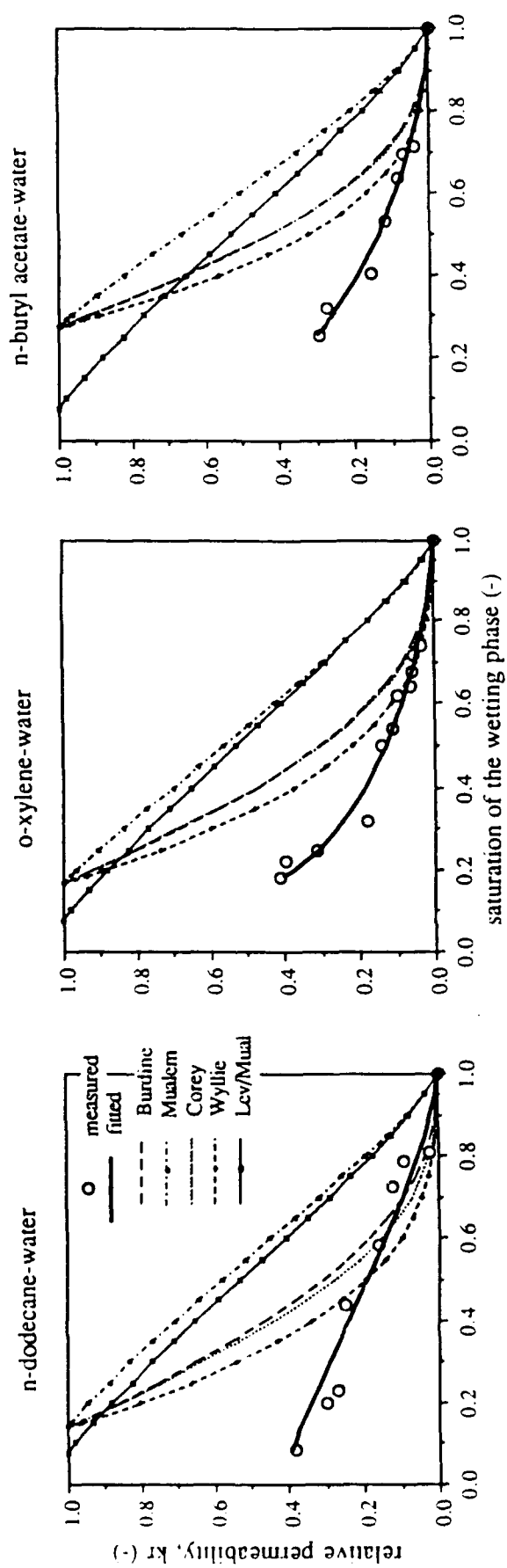


Figure 3. Comparison of measured and estimated relative permeability-saturation relationships for the organic liquid phase.



IV. METHODOLOGY

To investigate the effect of estimated transport relationships on predictions of multiphase flow, a simple, one-dimensional scenario is hypothesized. An aquifer is confined and at atmospheric pressure. It has no natural pressure gradient and is initially saturated with water. An underground storage tank filled with organic liquid located at $x=0$ springs a leak at time $t=0$ (Fig. 4). The head in the tank is sufficient to displace all the water at the upstream boundary. The tank has a sufficiently large circumference so that leakage into the aquifer does not significantly affect the height of liquid in the tank. The downstream boundary, $x=L$, is located at a sufficient distance so that the front of the organic liquid does not reach the boundary. The appropriate equations, initial and boundary conditions for such a situation are:

$$\frac{\partial}{\partial x} \left[\frac{\rho_w k k_{rw}}{\mu_w} \left(\frac{\partial P_w}{\partial x} \right) \right] = \frac{\partial (n \rho_w S_w)}{\partial t} \quad [8]$$

$$\frac{\partial}{\partial x} \left[\frac{\rho_n k k_{rn}}{\mu_n} \left(\frac{\partial P_n}{\partial x} \right) \right] = \frac{\partial (n \rho_n S_n)}{\partial t} \quad [9]$$

$$\begin{aligned} \text{at } t = 0, \quad P_w &= 0 \\ P_n &= 0 \quad \text{for all } x; \end{aligned}$$

$$\begin{aligned} \text{at } t > 0 \quad P_w &= 0 \quad \text{at } x = 0 \\ P_n &= 50 \text{ cm} \end{aligned}$$

$$\begin{aligned} P_w &= 0 \quad \text{at } x = L \\ \frac{\partial P_c}{\partial x} &= 0 \end{aligned}$$

Three organic liquid-water pairs were employed in the simulations. Their properties are summarized in Table 2. n-Dodecane and o-xylene are components of JP-4 (Smith et al., 1981) and clean JP-4 has a similar density, viscosity, and liquid-liquid interfacial tension to n-dodecane (Sittig, 1985). The porous medium was similar to the sand at Tyndall AFB. The similarity in properties is shown in Figure 5 (Demond, 1989; Demond and Roberts, 1991a). It was a clean, fine to medium sand from a contaminated aquifer located in Borden, Ontario (Mackay et al., 1986). The porosity and intrinsic permeability of the sand

is given in Table 3. More details about the liquids and the sand can be found in Demond and Roberts (1991a and 1991b).

Figure 4. Scenario for simulations of organic liquid migration.

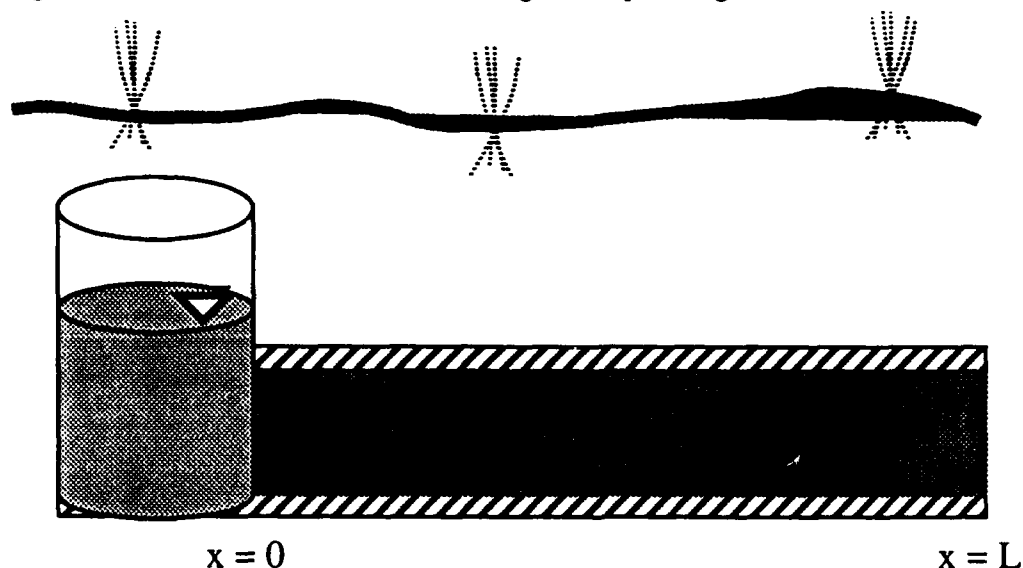


TABLE 2. Properties of Liquids Used in Simulations

Liquid	ρ (g/cm ³)	μ (cp)	$\gamma_{\text{ORG}/\text{H}_2\text{O}}$ (dynes/cm)	Contact angle on glass* (°)
n-dodecane	0.7487	1.508	52.8 [25°C] [2]	17
o-xylene	0.8801	0.810	36.1 [3]	25
n-butyl acetate	0.8764 [25°C]	0.7375	14.5 [25°C] [1]	59
water	0.9882	1.002	NA	NA

*The second liquid is water. The angle is reported as measured through the aqueous phase. All contact angle measurements are from Demond (1988). All values are at 20°C unless otherwise indicated. All other data are from Riddick et al. (1986) unless indicated by [number]: [1] Donahue and Bartell (1952); [2] Johnson and Dettre (1966); [3] Young and Harkins (1928).

Figure 5. Comparison of properties of Borden and Tyndall sands.

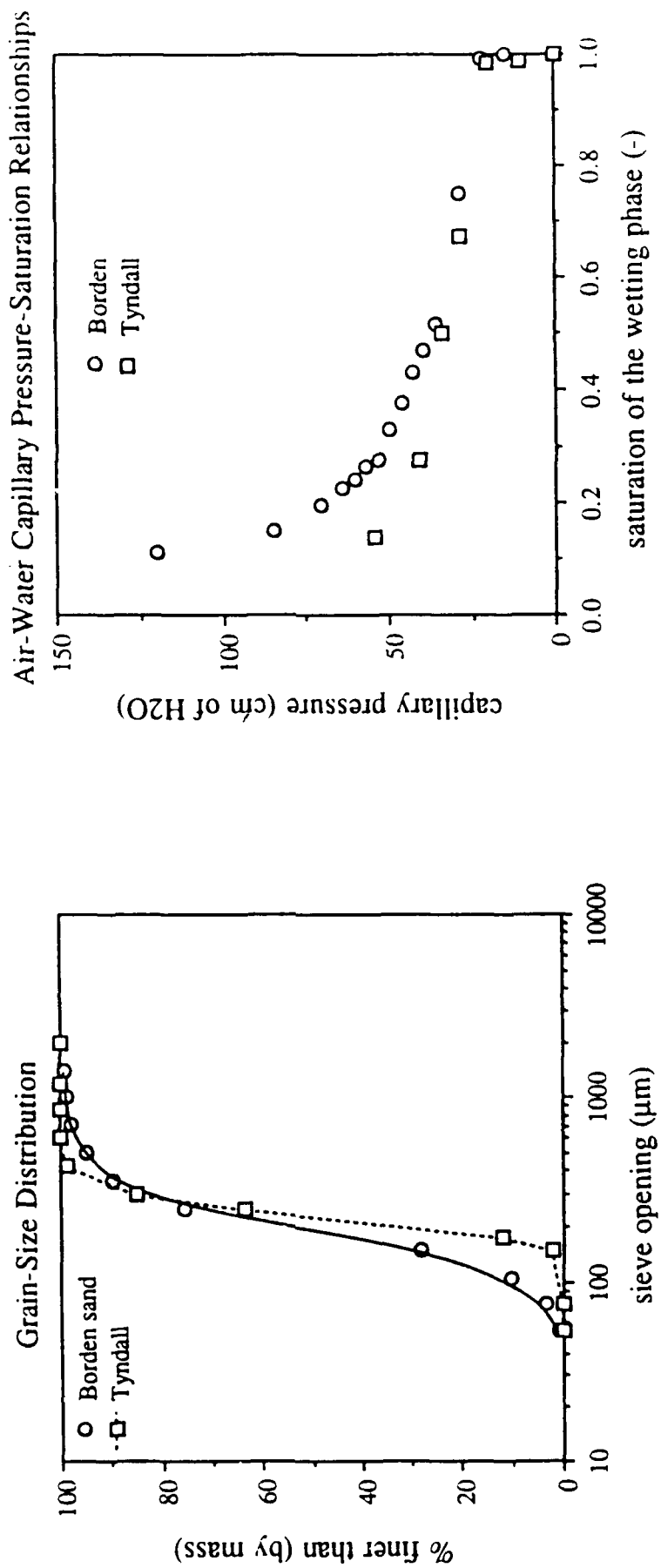


TABLE 3. Properties of Borden Sand

porosity (-)	intrinsic permeability (cm ²)
0.33	8.4x10 ⁻⁸

The measurements of capillary pressure and relative permeability were obtained from Demond and Roberts (1991a) and (1991b) respectively. The capillary pressure data were then fit with van Genuchten's function (van Genuchten, 1980):

$$S_e = \left[\frac{1}{1 + (\alpha P_c)^n} \right]^{1-1/n} \quad [10]$$

where n and α = fitting constants,

using the statistics package SYSTAT (SYSTAT, Inc. Evanston, IL) to determine n and α through nonlinear regression.

Several functional forms for the relative permeability data were tried. The most successful were the following equations, derived from the equations given in Table 1:

$$k_{rw} = S_e^a [1 - (1 - S_e^{1/b})^b] \quad [11]$$

and

$$k_m = c (1 - S_e)^d (1 - S_e^{1/e})^e \quad [12]$$

where a , b , c , d , and e = fitting constants.

The constants, a , b , c , d , and e were also determined using SYSTAT.

For the two-phase flow simulations, the original intention was to use SWANFLOW (GeoTrans, 1989), a commercially available finite difference code. Two versions of this code, the original, Version 1.0, and the most recent, Version 3.0, were purchased. Unfortunately, both contained problems and the documentation accompanying the codes was sufficiently incomplete that it was difficult to resolve the problems. Eventually, this

code was abandoned in favor of a code developed by Reeves and Abriola (Reeves and Abriola, 1988; Reeves, 1991). The selected code uses finite element approximation for the spatial derivatives. Linear basis functions are employed and weighting is based on the streamline-upwind Petrov-Galerkin technique. The time derivatives are approximated using finite differences and the resulting matrices are solved using successive substitution or full Newton-Raphson should initial convergence prove unsatisfactory. The code was executed on the Apollo network of workstations operated by CAEN (Computer Aided Engineering Network) at the University of Michigan. The nodal spacing employed in the simulations was 0.75 cm and the length of the domain was 45 cm. Due to computational constraints, the time span of the simulations was limited to 1000 secs.

Three series of simulations were performed:

- 1) using measured data for both capillary pressure and relative permeability, parameterized using Eqn. [10], and Eqns. [11] and [12], respectively,
- 2) using measured capillary pressure relationships, but estimated relative permeabilities based on the correlations in Table 1, and
- 3) using estimates for both relationships, with capillary pressure relationships for organic liquid-water pairs based on air-water data (parameterized using Eqn. [10]) and Eqn [6], and with relative permeability relationships given by Mualem's correlation (Table 1). This combination was selected because it appears to be the most popular in simulations of subsurface organic liquid flow.

V. RESULTS AND DISCUSSION

The ability of van Genuchten's function (Eqn. 10) to fit the capillary pressure data is shown in Figure 6. The ability of Eqns. [11] and [12] to fit the relative permeability data is shown in Figures 2 and 3. These figures demonstrate that these functional forms are suitable for the parameterization of organic liquid-water transport relationships. The parameters for capillary pressure and relative permeability relationships are given in Tables 4 and 5, respectively.

Figure 7 shows a sample simulation using measured data, illustrating the progression of the organic liquid front over time. The fronts at two times, 495 seconds and 1000 seconds, are shown again in Figures 8 and 9, respectively; but here, these fronts are compared to those generated using estimated relationships. The comparison shown in Figures 8 and 9 reveals

a significant discrepancy between the predicted location of the front using measured relationships and those using estimated relationships. Most of this discrepancy probably stems from the poor approximation of the relative permeability to the organic liquid (Figures 1-3). The discrepancy grows if both the capillary pressure and relative permeability relationships are predicted. The higher saturations at $x=0$ produced in the simulations for the combination of estimated capillary pressure and estimated relative permeability (Series 3) result from the fact that Leverett's function predicts a lower residual saturation than measured (Figure 1). Comparing Figures 8 and 9 shows that the discrepancy increases with time. Based on this observation at short times, one would expect even larger differences at time scales typical of field problems.

Table 6 lists estimates of the total volume of organic liquid infiltrated per unit area, corresponding to the displacement fronts shown in Figures 8 and 9. The error in the estimate of the amount of organic liquid infiltrated is similar for all four methods for calculating relative permeability, ranging from about 10 to 12% at 495 seconds. If the capillary pressure is also estimated, the error increases to about 25%. At longer times, the error may be greater. At 1000 seconds, the error resulting from estimation of both capillary pressure and relative permeability is on the order of 30%.

The simulations shown here are all of the drainage process. Since previous work shows that the discrepancies between measured and estimated imbibition relationships are even greater, one would expect the errors to be larger in simulations of the imbibition process.

VI. CONCLUSIONS

The results presented here suggest that if the capillary pressure and relative permeability are estimated using techniques commonly employed, the location of the displacement front and the volume of organic liquid infiltrated may be grossly overpredicted. In the field, these errors may translate into an inability to locate the boundaries of a jet fuel spill and an overestimation of the amount of fuel that is present in the subsurface. The results point to the need to use measured relationships rather than estimated in simulations of two-phase flow. Yet, it is impractical to measure these relationships in many circumstances. Consequently, attention must be devoted to the development of more accurate estimation techniques for capillary pressure and relative permeability for organic liquids in the subsurface.

Figure 6. Fit of van Genuchten's function to drainage capillary pressure-saturation relationships.

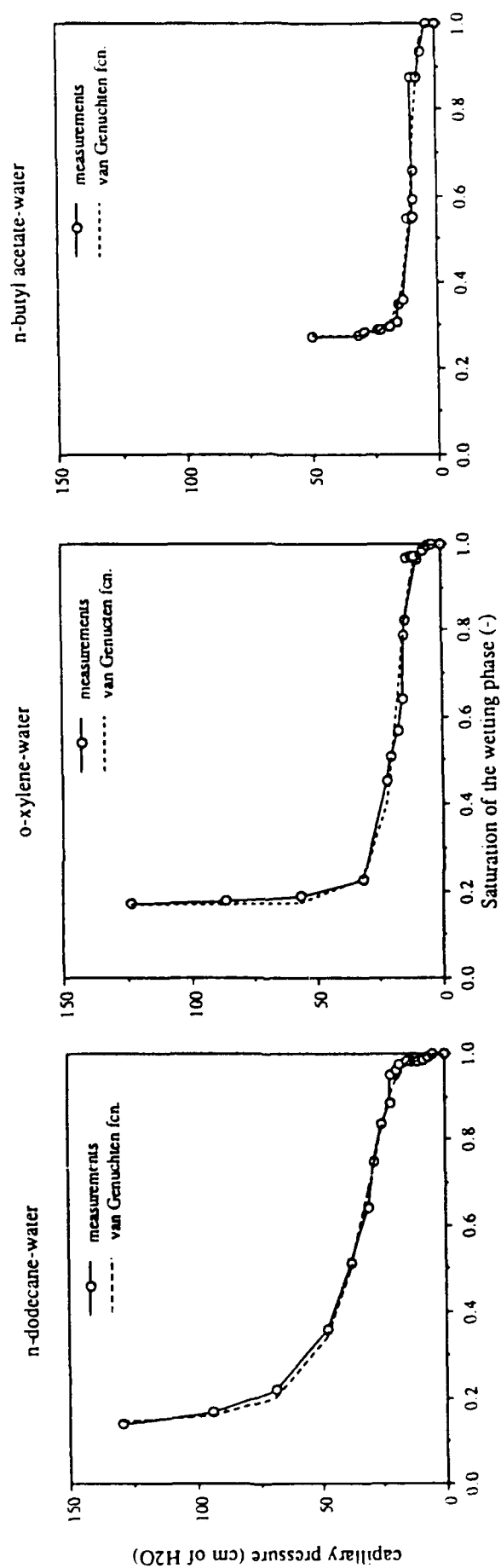


Figure 7. Location of organic liquid front as a function of time.

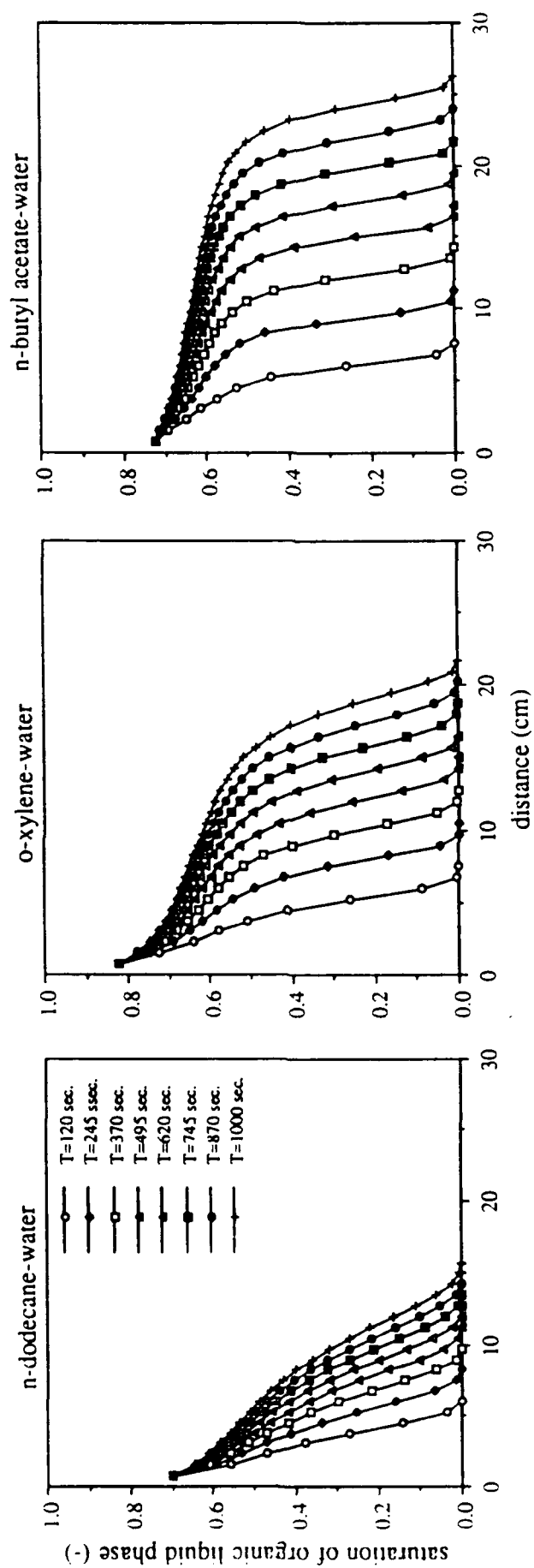


Figure 8. Displacement fronts at 495 seconds simulated using measured and estimated transport relationships.

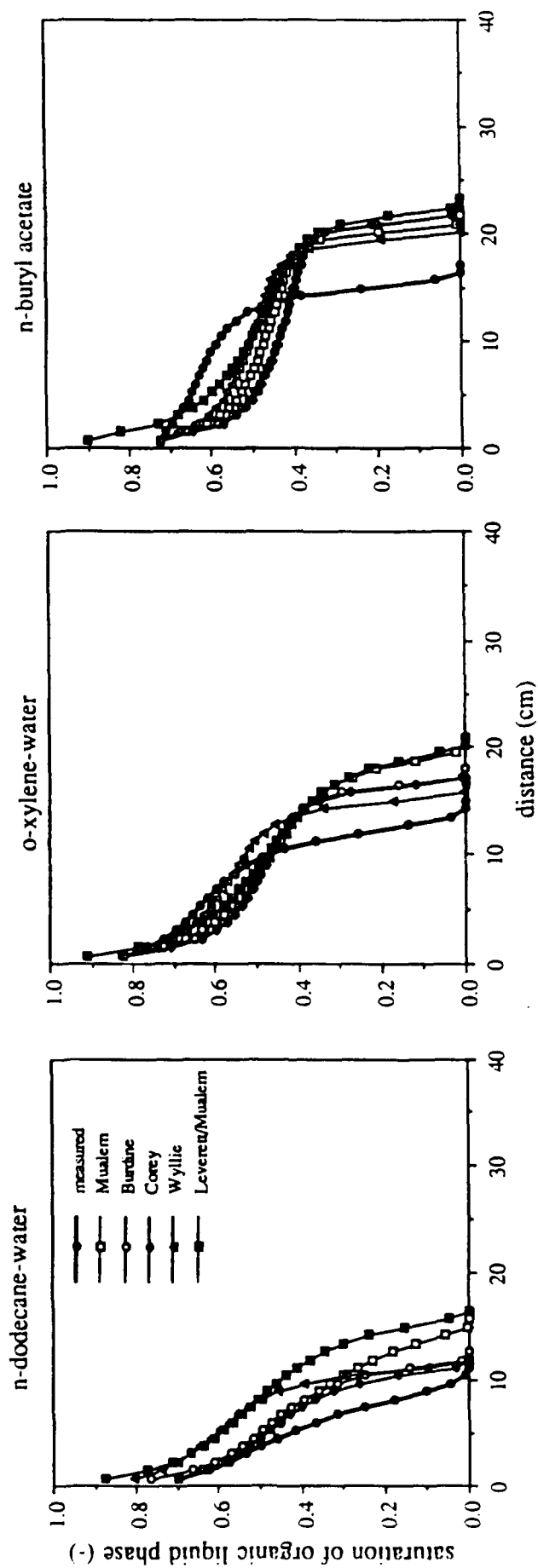


Figure 9. Displacement fronts at 1000 seconds simulated using measured and estimated transport relationships.

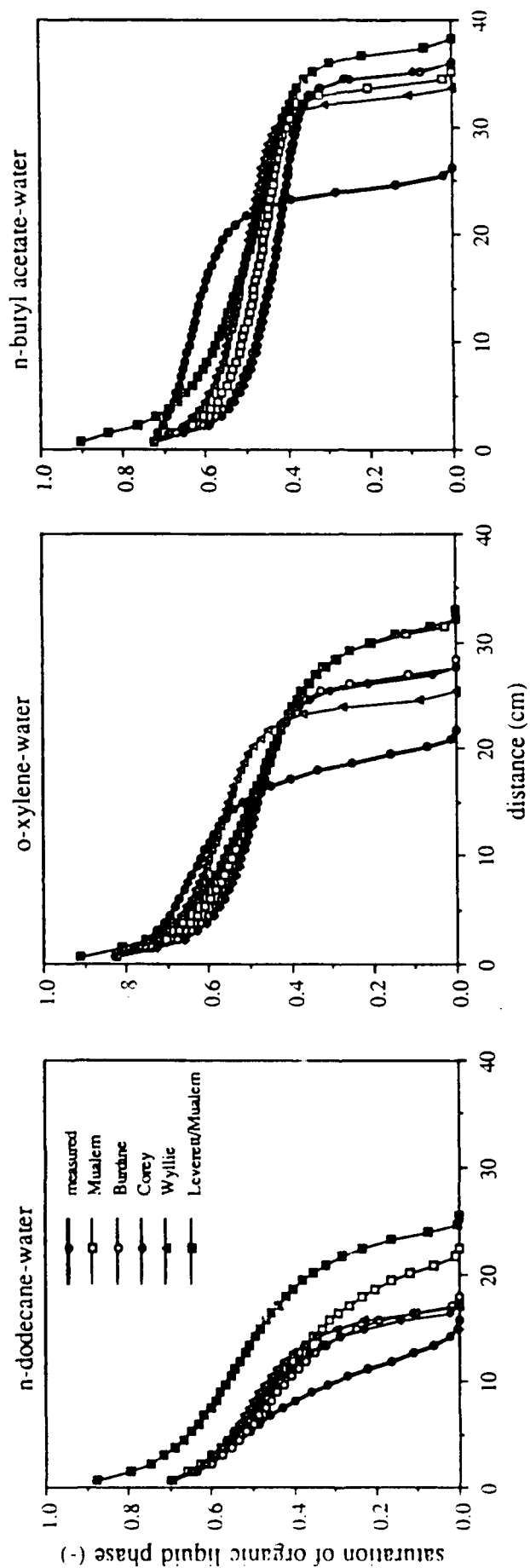


TABLE 4. Parameters for Measured Capillary Pressure-Saturation Relationships

	Drainage			Imbibition		
	α	n	S_{wr}	α	n	S_{nr}
air-water	0.030	3.969	0.073	0.042	3.769	0.120
n-dodecane-water	0.038	3.882	0.139	0.075	4.436	0.151
o-xylene-water	0.057	5.446	0.170	0.098	4.374	0.193
n-butyl acetate-water	0.106	5.318	0.273	0.189	3.185	0.702

TABLE 5. Parameters for Measured Relative Permeability-Saturation Relationships

AQUEOUS PHASE			
	Drainage		Imbibition
	a	b	a b
n-dodecane-water	1.283	0.578	0.567 0.458
o-xylene-water	2.331	1.208	2.292 1.200
n-butyl acetate-water	1.000	1.000	1.100 0.737

ORGANIC LIQUID PHASE					
	Drainage			Imbibition	
	c	d	e	c	d e
n-dodecane-water	0.390	1.000	0.500	0.390	1.600 0.550
o-xylene-water	0.425	0.159	1.479	0.425	3.738 0.806
n-butyl acetate-water	0.300	0.672	1.130	0.300	3.818 0.055

TABLE 6. Comparison of Volumes infiltrated Per Unit Area of Aquifer Using Measured and Estimated Transport Relationships

Time = 495 sec	Measured properties	Burdine (1953)	Muallem (1976)	Corey (1954)	Wyllie (1962)	Leverett (1941) Muallem (1976)
n-dodecane-water	1.21	1.64	1.78	1.49	1.92	2.42
o-xylene-water	2.30	2.59	2.82	2.55	2.65	2.94
n-butyl acetate-water	2.83	3.01	3.10	3.00	3.18	3.55

Time = 1000 sec	Measured properties	Burdine (1953)	Muallem (1976)	Corey (1954)	Wyllie (1962)	Leverett (1941) Muallem (1976)
n-dodecane-water	1.83	2.57	2.78	2.32	3.06	3.94
o-xylene-water	3.72	4.31	4.79	4.24	4.47	4.95
n-butyl acetate-water	4.73	5.03	5.24	5.01	5.44	6.13

Units are cm^3/cm^2

VII. RECOMMENDATIONS

The research presented in this report shows that the location of the organic liquid displacement front and the amount of organic liquid infiltrated may be significantly overpredicted if estimated, rather than measured, transport relationships are used in the transport simulations. This error may translate into large overestimates of quantities of jet fuel located in the subsurface at a given spill site, or an inability to locate the boundaries of a subsurface spill. This conclusion results in two recommendations:

- 1) Caution must be exercised in basing assessment and clean-up strategies on transport simulations which employ estimated two-phase flow transport relationships. An analysis which examines the consequences of estimates being significantly in error should be part of any assessment program.
- 2) To make simulations of two-phase flow more useful in subsurface remediation, research must be devoted towards developing more accurate methods for the estimation of capillary pressure and relative permeability relationships. The research should address not only the reasons why the common methods for drainage fail to do an adequate job, but also the lack of accepted means for estimating imbibition relationships.

VIII. REFERENCES

- Abriola, L.M., and G.F. Pinder, 1985, A multiphase approach to the modeling of porous media contamination by organic compounds, 1. Equation development, *Water Res. Res.*, 21(1), 11-18.
- Burdine, N.T., 1953, Relative permeability calculations from pore size distribution data, *Trans. AIME*, 198, 71-78.
- Cary, J.W., J.F. McBride, and C.S. Simmons, 1989, Observations of water and oil infiltration into soil: some simulation challenges, *Water Res. Res.*, 25(1):73-80.
- Corapcioglu, M.Y., and A.L. Baehr, 1987, A compositional multiphase model for groundwater contamination by petroleum products, 1. Theoretical considerations, *Water Res. Res.*, 23(1), 191-200.
- Corey, A.T., 1954, The interrelation between gas and oil relative permeabilities, *Producers Monthly*, 19(1), 38-41.
- Demon, A.H., 1988, Capillarity in Two-Phase Liquid Flow of Organic Contaminants in Groundwater, Ph.D. thesis, Stanford University, Stanford CA, 1988. 211pp.
- Demon, A.H., 1989, Prediction of the Capillary Pressure-Saturation Relationships for Aquifers Contaminated with Jet Fuels, Final Report to Air Force Office of Scientific Research, Contract No. F49620-88-C-0053, August 1989.
- Demon, A.H., and P.V. Roberts, 1991a, Effect of interfacial forces on two-phase capillary pressure relationships, *Water Res. Res.*, 27(3):423-437.
- Demon, A.H., and P.V. Roberts, 1991b, Estimation of Two-Phase Relative Permeability Relationships for Organic Liquid Contaminants, submitted for publication to *Water Res. Res.*
- Donahue, D.J., and F.E. Bartell, 1952, The boundary tension at water-organic liquid interfaces, *J. Phys. Chem.*, 56, 480-484.

- Faust, C.R., 1985, Transport of immiscible fluids within and below the unsaturated zone, a numerical model, *Water Res. Res.*, 21(4), 587-596.
- GeoTrans, 1989, SWANFLOW: Simultaneous Water, Air, and Nonaqueous Phase Flow, Version 3.0, GeoTrans, Herndon, VA.
- Johnson, R.E., Jr., and R.H. Dettre, 1966, The wettability of low energy liquid surfaces, *J. Colloid Interf. Sci.*, 21, 610-622.
- Kuppusamy, T., J. Sheng, J.C. Parker, and R.J. Lenhard, 1987, Finite-element analysis of multiphase immiscible flow through soils, *Water Res. Res.*, 23(4), 625-631.
- Lenhard, R. J., and J. C. Parker, 1987, Measurement and prediction of saturation-pressure relationships in three-phase porous media systems, *J. Contaminant Hydrol.* 1, 407-424.
- Leverett, M. C., 1941, Capillary behavior in porous solids, *Trans. AIME*, 142, 152-169.
- Mackay, D.M., D.L. Freyberg, and P.V. Roberts, 1986, A natural gradient experiment on solute transport in a sand aquifer, 1. Approach and overview of plume movement, *Water Res. Res.*, 22(13), 2017-2029.
- Mualem, Y., 1976, A new model for predicting the hydraulic conductivity of unsaturated porous media, *Water Res. Res.*, 12(3), 513-522.
- Osborne, M., and J. Sykes, 1986, Numerical modeling of immiscible organic transport at the Hyde Park landfill, *Water Res. Res.*, 22(1), 25-33.
- Parker, J.C., R.J. Lenhard, and T. Kuppusamy, 1987, A parametric model for constitutive properties governing multiphase flow in porous media, *Water Res. Res.*, 23, 618-624.
- Pinder, G.F., and L.M. Abriola, 1986, On the simulation of nonaqueous phase organic compounds in the subsurface, *Water Res. Res.*, 22(9), 109s-119s.
- Reeves, H., and L.M. Abriola, 1988, A decoupled approach to the simulation of flow and transport of nonaqueous organic phase contaminants through porous media, in *Modeling*

- Surface and Subsurface Flows, Vol. 1.*, edited by M.A. Celia and L.A. Ferrand, pp. 147-152, Elsevier, New York.
- Reeves, H., 1991, Volatilization and Vapor Phase Transport of Organic Contaminants in the Subsurface, Ph.D. thesis, Univ. of Michigan, Ann Arbor, MI.
- Riddick, J.A., W.B. Bunger, and T.K. Sakano, 1986, *Organic Solvents: Physical Properties and Methods of Purification*, 4th ed., Wiley, New York.
- Schwille, F., 1984, Migration of organic fluids immiscible with water in the unsaturated zone, in *Pollutants in Porous Media: The Unsaturated Zone Between Soil Surface and Groundwater*, edited by B. Yaron, G. Dagan and T. Goldschmid, pp. 27-48, Springer-Verlag, New York.
- Sittig, M., 1985, *Handbook of Toxic and Hazardous Chemicals and Carcinogens*, 2nd ed., Noyes Pub., Park Ridge, NJ.
- Smith, J.H., J.C. Harper, and H. Jayber, 1981, Analysis and Environmental Fate of Air Force Distillate and High Density Fuels, Rpt. No. ESL-TR-81-54, Air Force Engineering and Services Center, Panama City, FL. 144pp.
- van Genuchten, M.Th., 1980, A closed form solution for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Amer. J.*, 44, 892-898.
- Wyllie, M.R.J., 1962, Relative permeability, in *Petroleum Production Handbook, Vol. II: Reservoir Engineering*, edited by T.C. Frick and R.W. Taylor, pp. 25.1-25.14, McGraw-Hill, New York.
- Young, T.F., and W.D. Harkins, 1928, Interfacial tension of solid-liquid and liquid-liquid interfaces, in *International Tables, Vol IV*, edited by E.W. Washburn, pp. 436-439, McGraw-Hill, New York.

IX. TABLES AND FIGURES

- Table 1. Correlations for Estimation of Drainage Relative Permeabilities
- Table 2. Properties of Liquids Used in Simulations
- Table 3. Properties of Borden Sand
- Table 4. Parameters for Capillary Pressure-Saturation Relationships
- Table 5. Parameters for Relative Permeability-Saturation Relationships
- Table 6. Comparison of Volumes Infiltrated Per Unit Area of Aquifer Using Measured and Estimated Transport Relationships
-
- Figure 1. Comparison of measured and estimated capillary pressure-saturation relationships.
- Figure 2. Comparison of measured and estimated relative permeability-saturation relationships for the aqueous phase.
- Figure 3. Comparison of measured and estimated relative permeability-saturation relationships for the organic liquid phase.
- Figure 4. Scenario for simulations of organic liquid migration.
- Figure 5. Comparison of properties of Borden and Tyndall sands.
- Figure 6. Fit of van Genuchten's function to drainage capillary pressure-saturation relationships.
- Figure 7. Location of organic liquid front as a function of time.
- Figure 8. Displacement fronts at 495 seconds simulated using measured and estimated transport relationships.
- Figure 9. Displacement fronts at 1000 seconds simulated using measured and estimated transport relationships.

1989 USAF-UES MINI GRANT PROGRAM

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Conducted by the
University Energy Systems, Inc.

LABORATORY INVESTIGATIONS OF SUBSURFACE
CONTAMINANT SORPTION SYSTEMS

Prepared by:	Kirk Hatfield, Ph.D. and Joe Ziegler
Academic Rank:	Assistant Professor / Graduate Student
Department and	Civil Engineering
University	University of Florida
Research Location:	University of Florida
USAF Researcher:	Thomas B. Stauffer, Ph.D.
Date:	28 Feb 91
Contract No.:	F49620-88-C-0053/SB5881-0378

**LABORATORY INVESTIGATIONS OF SUBSURFACE
CONTAMINANT SORPTION SYSTEMS**

by

Kirk Hatfield and Joe Ziegler

ABSTRACT

Enhancing the sorptive capacity of natural aquifer materials is a relatively new approach to bringing about aquifer remediation and protection. Previous investigations had focused on the application of cationic surfactants to enhance sorption; however, the utility of any cationic surfactant is constrained by the cation exchange capacity of the target aquifer and the biocidal characteristics of the surfactant. An alternative sorbent could be an innocuous form of a sorptive nonaqueous phase liquid (SNAPL). The first half of this report presents evidence of significant enhanced sorption in soil columns possessing a nontoxic SNAPL residual. Breakthrough curves (BTCs) of pentafluorobenzoic acid (PFBA), benzene, xylene, PCE, toluene, and 1-methylnaphthalene were measured during aqueous elution through soil columns totally saturated with water and through soil columns containing a residual saturation of decane. Three conceptual models were used to simulate premature breakthrough, asymmetry, and tailing in the BTCs caused by nonequilibrium sorption and rate limited mass transfer between decane and aqueous phases. The second half of the report is devoted to the development and application of a new concept, the 'effective retardation' which can be used in subsurface sorption system design and in the prediction of system performance.

ACKNOWLEDGMENTS

We would like to acknowledge the efforts of those who facilitated in the completion of this work. Special gratitude is extended to Drs. Tom Stauffer and David Burris for their guidance, insight, and support. Chris Antworth is thanked for his technical assistance and advice. Finally, we

wish to express our appreciation to the Air Force Systems Command, and the Air Force Engineering Services Center for sponsoring this research.

1.0 INTRODUCTION

Groundwater contamination resulting from fuels and oils is particularly nefarious because of the widespread use of petroleum based products and the health risks associated with the hydrophobic organic components. Natural geologic materials generally have a low retentive capacity for neutral organic compounds (Schwarzenbach and Westall 1981; Banerjee et al. 1985; Bouchard et al. 1988a; and Stauffer et al. 1989). Little or no natural retention reduces local constituent residence times and decreases the opportunity for biotic and abiotic processes to naturally attenuate contaminants.

Enhanced subsurface sorption has been suggested as an innovative groundwater remediation technique (Burris and Antworth 1990 and Hatfield et al. 1991) which could be used alone or in conjunction with more traditional remediation technologies. Sorption zones could be created within the aquifer. Enhancing the sorption capacity of the porous matrix would retard contaminant movement and concentrate pollutants within localized regions of the aquifer. One or more of these sorption zones would constitute a subsurface sorption system (SSS). Fig 1. illustrates the conceptual application of a SSS to intercept a plume of contaminants released into a surficial aquifer.

The use of cationic surfactants to increase the retardation factors in geologic materials has been the focus of several recent investigations. Lee et al. (1989) produced a 200 fold increase in sorption coefficients for a subsurface soil treated with an organic cation (hexadecyltrimethylammonium). Burris and Antworth (1990) modified aquifer material with cationic surfactants and subsequently increased the sorption coefficients for common groundwater contaminants by three orders of magnitude. Effective use of cationic surfactants is highly dependent upon the cation exchange capacity (CEC) of the aquifer material. Many aquifer materials have low CEC's, thus limiting the potential application of cationic surfactants. Many of the cationic surfactants of interest have

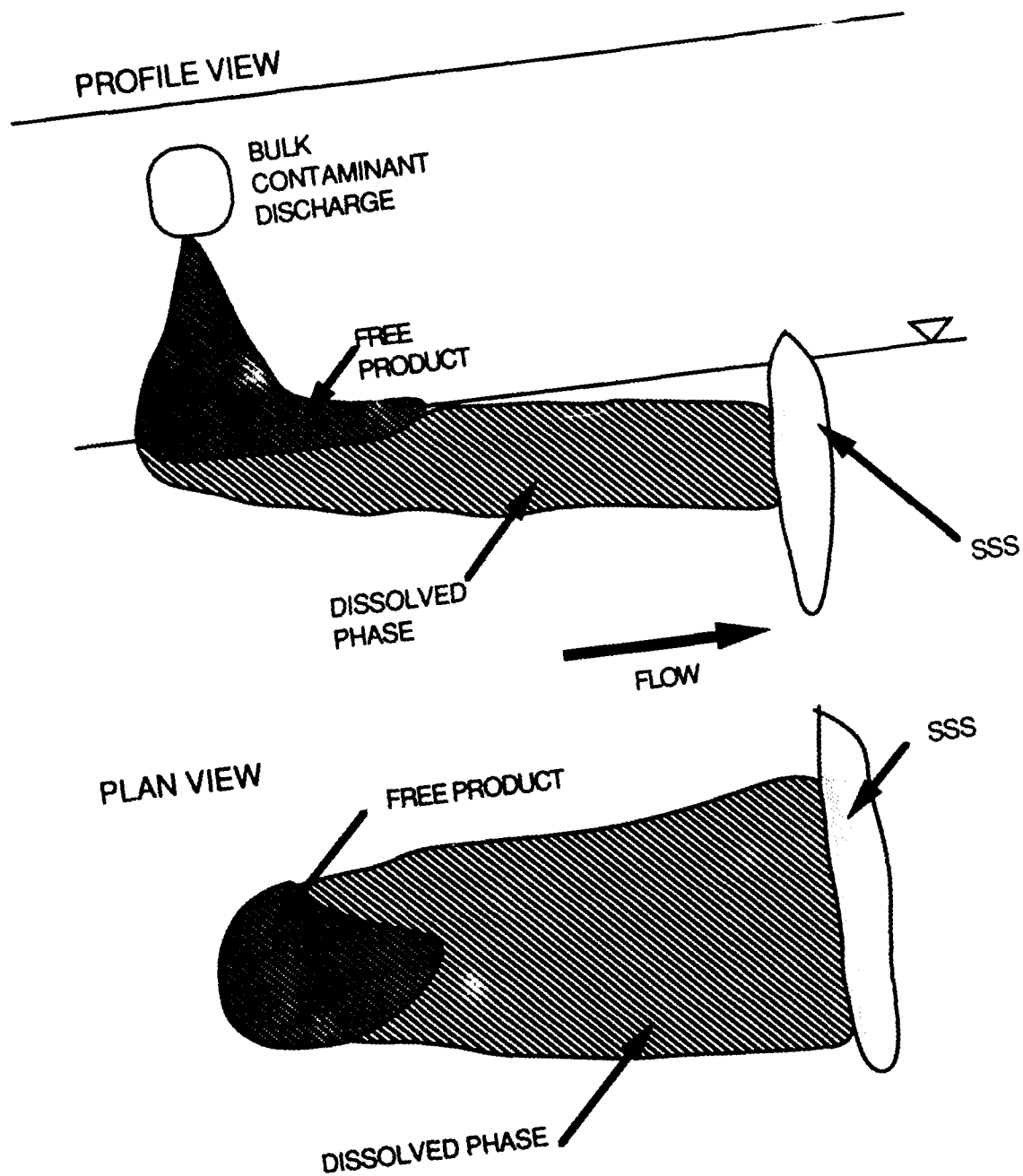


Fig. 1 Concept 1 Illustration of Subsurface Sorption System.

biocidal activity that may limit the ability to couple a cationic surfactant with biodegradation. Clearly, other methods of enhancing sorption need to be examined.

A suggested alternative to organic cations would be a nontoxic, nonaqueous phase liquid (NAPL). The sorptive capacity of subsurface materials treated with a sorbing NAPL (SNAPL) would depend upon the residual SNAPL saturation and the contaminant SNAPL-water partition coefficient. The residual saturation of SNAPL is not, however, limited by the available CEC of the soil, as is the case with cationic surfactants.

The first half of this report examines the transport and retention of several dissolved organics in a soil partially saturated with a SNAPL (decane). Miscible displacement experiments were performed with water saturated soil columns and with other soil columns saturated with water and residual decane. Organic tracers used in the study were pentfluorobenzoic acid (PFBA), benzene, xylene, PCE, toluene, and 1-methylnaphthalene. Three existing transport models were used to simulate the premature solute breakthrough, asymmetry, and tailing in the BTCs caused by physical and sorption related nonequilibrium. Nonequilibrium modeling is of interest, because the conditions giving rise to nonequilibrium sorption may result in premature breakthrough of contaminants through the sorption system. It is therefore necessary to investigate conditions causing nonequilibrium and to identify suitable models to predict premature failure of the sorption system. The second half of the report is devoted to predicting the performance of a SSS created to enhance the contaminant retardation capabilities of an aquifer. The concept of 'effective retardation' is introduced as an aid to predict system performance and to facilitate system design. An analysis develops the necessary equations for estimating the performance of a cylindrical sorption system, while an application demonstrates the utility of the effective retardation concept.

2.0 OBJECTIVES

The goal of this research effort was to develop a knowledge base useful in the design of sorption systems and the prediction of system performance in the field. The initial objectives specified; 1) a verifi-

cation effort for a 3-dimensional immiscible flow model, 2) bench-scale studies to generate data for model verification, and 3) a series of miscible displacement experiments with xylene and 1-methylnaphthalene, which would provide data needed to predict sorptive system capacity and performance.

Originally, it was presumed that direct injection of the sorbing NAPL (SNAPL) was the best way to install a sorption zone. Direct injection, however, is not likely to create a uniform distribution of SNAPL in the porous media. Without knowing the spatial distribution of SNAPL, it would be difficult to predict sorption system performance; this is because distribution of SNAPL affects sorption capacity and contaminant velocities. It was then surmised that a verified 3-dimensional immiscible flow model would be needed to predict residual SNAPL saturations and pore velocities throughout the sorption zone; and it was this need, that justified the research objectives leading to the verification of an immiscible flow model. Early in the project, however, a review of ground modification techniques, revealed that there are desirable alternatives to direct injection. One alternative was to use jet grouting to create sorption zones of desired dimensions. Jet grouting removes the subsurface soils without surface excavation. Subsurface soil are then replaced with a substitute material. Jet grouting offers control over sorption zone dimensions, permeability, and the residual SNAPL distribution. Clearly, with an emplacement technique such as jet grouting, model predictions of residual SNAPL saturations are no longer necessary; hence, there was no longer a need to verify an immiscible flow model. On-the-other-hand, predictions of pore water velocities inside and around the sorption system are still needed to predict system performance.

Given the availability of an emplacement technique that nullifies initial justifications for verifying an immiscible flow model, project objectives were amended in consultation with the project officer. The amended objectives were formulated with careful attention given to preserve the original goal of this research effort; that is, to develop a knowledge base useful in the design of sorption systems and the prediction of system performance in the field. The amended objectives also reflect

additional miscible displacement experiments. Early research efforts were quite successful in column breakthrough experiments. After consultation with the project officer, it was agreed that additional experiments were necessary and in the best interest of meeting the project goal. The finalized objectives of this research effort were:

- 1) to perform column breakthrough experiments on Xylene, 1-Methylnaphthalene, Benzene, Toluene, and PCE which deliver data needed to predict sorption system capacity and performance for these contaminants,
- 2) to develop general analytical expressions which can define the velocity field and predict the performance of subsurface sorption systems of cylindrical geometry.

3.0 MISCIBLE DISPLACEMENT EXPERIMENTS

3.1 THEORY

3.1.1 Model 1. Two Region Nonequilibrium Model

Premature contaminant breakthrough, asymmetry, and tailing in BTCs for hydrophobic contaminants may be due to large heterogeneities in pore water velocities. The velocity variations create a physical nonequilibrium condition that affects both sorbing and nonsorbing solutes. The two region nonequilibrium model, first proposed by Coats and Smith (1964) for nonreactive solutes, presumes that the complex saturated media may be modeled as a two region domain. In the first region the fluid is mobile, which is where convective dispersive solute transport occurs. In the second region, the fluid is relatively stagnate or immobile. The 'mobile' and 'immobile' regions are connected; but, rate-limited diffusion controls solute exchange between the mobile and immobile regions. Van Genuchten and Wierenga (1976) extended the nonreactive contaminant model to incorporate linear sorption. The governing equation describing transport in the mobile region is

$$\begin{aligned} \theta_m \frac{\partial C_m}{\partial t} + f \rho \frac{\partial S_m}{\partial t} + \theta_{im} \frac{\partial C_{im}}{\partial t} + (1 - f) \rho \frac{\partial S_{im}}{\partial t} \\ = \theta_m D_m \frac{\partial^2 C_m}{\partial x^2} - \theta_m V_m \frac{\partial C_m}{\partial x} \end{aligned} \quad (1)$$

Transport in the immobile region is given by

$$\theta_{im} \frac{\partial C_{im}}{\partial t} + (1 - f) \rho \frac{\partial S_{im}}{\partial t} = k_1 (C_m - C_{im}) \quad (2)$$

in which C_m and S_{im} are respectively the dissolved and sorbed solute concentrations in the mobile region, and C_{im} and S_{im} are the corresponding dissolved and sorbed phase concentrations in the immobile region. The bulk density of the porous media is ρ . The total volumetric water content of the saturated media, θ , is the summation of volumetric water content for mobile and immobile regions, θ_m and θ_{im} , respectively. The mass fraction of sorption sites associated with the mobile fluid is f . Advective and dispersive transport components, that are pertinent to the mobile region, appear on the right side of Eq. 1 and are expressed in terms of V_m , the average mobile pore-water velocity and D_m , the apparent dispersion coefficient. Finally, the exchange of solutes between the mobile and immobile liquid regions is described using a first-order mass transfer coefficient, k_1 .

Sorption in both the mobile and immobile regions of the solid matrix is assumed to be linear, reversible and instantaneous. Therefore, in each region the sorbed contaminant concentration is

$$S_m = K_d C_m \quad (3a)$$

and

$$S_{im} = K_d C_{im} \quad (3b)$$

where K_d is the distribution coefficient for linear adsorption. The time rate of change of sorbed concentrations is

$$\frac{\partial S_m}{\partial t} = K_d \frac{\partial C_m}{\partial t} \quad (4a)$$

and

$$\frac{\partial S_{im}}{\partial t} = K_d \frac{\partial C_{im}}{\partial t} \quad (4b)$$

The results presented herein are from one dimensional miscible displacement experiments, for which a pulse of contaminant is introduced over a time period t_0 . In experiments, where transport related nonequilibrium was considered, system behavior can be approximated by solving Eqs. 1 and 2 under the following initial and boundary conditions

$$C_m(x, 0) = C_{im}(x, 0) = 0 \quad (5a)$$

$$S_m(x, 0) = S_{im}(x, 0) = 0 \quad (5b)$$

$$-D_m \frac{\partial C_m}{\partial x} + V_m C_m \Big|_{x=0} = V_m C_0 \quad 0 < t \leq t_0 \quad (5c)$$

$$-D_m \frac{\partial C_m}{\partial x} + V_m C_m \Big|_{x=0} = 0 \quad t > t_0 \quad (5d)$$

$$\frac{\partial C_m}{\partial x} \Big|_{x=L} = 0 \quad t > 0 \quad (5e)$$

For miscible displacement experiments conducted with soil columns of length, L , a convenient dimensionless expression for the governing transport equations is the following

$$\beta R \frac{\partial C_1}{\partial T} + (1 - \beta) R \frac{\partial C_2}{\partial T} = \frac{1}{P} \frac{\partial^2 C_1}{\partial X^2} - \frac{\partial C_1}{\partial X} \quad (6)$$

for the mobile region and

$$(1 - \beta) R \frac{\partial C_2}{\partial T} = \omega (C_1 - C_2) \quad (7)$$

for the immobile region, where the dimensionless variables are defined

$$C_1 = C_m/C_o \quad C_2 = C_{im}/C_o \quad (8)$$

$$T = \frac{V_m \theta_m t}{L \theta} \quad (9)$$

$$X = \frac{x}{L} \quad (10)$$

$$P = \frac{V_m L}{D_m} \quad (11)$$

$$R = 1 + \frac{\rho K_d}{\theta} \quad (12)$$

$$\beta = \frac{\theta_m + f \rho K_d}{\theta + \rho K_d} \quad (13)$$

$$\omega = \frac{k_1 L}{\theta_m V_m} \quad (14)$$

Eqs. 6 and 7 were derived from Eqs. 1 and 2, using Eq. 4 to eliminate S_m and S_{im} ; and, then using Eqs. 8 through 14 to substitute dimensionless variables for dimensional variables. In a similar fashion, initial and boundary conditions were rewritten below in terms of the dimensionless variables

$$C_1(X, 0) = 0 \quad (15a)$$

$$C_2(X, 0) = 0 \quad (15b)$$

$$-\frac{1}{P} \frac{\partial C_1}{\partial X} + C_1 \Big|_{X=0} = 1 \quad 0 < T \leq T_o \quad (15c)$$

$$-\frac{1}{P} \frac{\partial C_1}{\partial X} + C_1 \Big|_{X=0} = 0 \quad T > T_o \quad (15d)$$

$$\frac{\partial C_1}{\partial X} \Big|_{X=1} = 0 \quad T > 0 \quad (15e)$$

3.1.2 Model 2. The Two Site Model

Early solute arrival, observed asymmetry, and tailing in BTCs may also occur as a result of heterogeneities in sorption sites. Differences in the rates at which solutes sorb creates a sorption related nonequilibrium condition affecting only sorbing solutes. The two site nonequilibrium model, proposed by Selim et al. (1976) and Cameron and Klute (1977), presumes that the complex saturated media may be treated as a two site sorption domain; sorption on type-1 sites is instantaneous, while sorption on type-2 sites is time-dependent. Letting F equal the mass fraction of sorption capacity associated with type-1 sites, the governing equations for the transport model are

$$\theta \frac{\partial C}{\partial t} + \rho \frac{\partial S_1}{\partial t} + \rho \frac{\partial S_2}{\partial t} = D\theta \frac{\partial^2 C}{\partial x^2} - v\theta \frac{\partial C}{\partial x} \quad (16)$$

and

$$\frac{\partial S_2}{\partial t} = k_2 [(1 - F) K_d C - S_2] \quad (17)$$

where C is the dissolved solute concentrations, S_1 is the type-1 sorbed solute concentration, and S_2 is the type-2 sorbed solute concentration. k_2 represents a first-order kinetic rate coefficient used to describe the rate limited sorption at type-2 sites. Since it is assumed that all the

fluid is moving, D and V are taken to be apparent dispersion coefficient and pore velocity of the total liquid phase, respectively.

Sorption for both sites is assumed to be linear and reversible; thus, at equilibrium sorption at both sites is given by

$$S_1 = FK_d C \quad (18)$$

and

$$S_2 = (1 - F) K_d C \quad (19)$$

The time dependent change in the sorbed solute concentrations at type-1 sites is

$$\frac{\partial S_1}{\partial t} = K_d \frac{\partial C}{\partial t} \quad (20)$$

whereas rate limited exchange of solutes at type-2 sites is expressed in Eq. 17.

BTCs are from one dimensional miscible displacement experiments, for which a pulse of contaminant is introduced over a time period t_o . The appropriate initial and boundary conditions under which Eqs. 17 and 18 should be solved are

$$C(x, 0) = 0 \quad (21a)$$

$$S_2(x, 0) = 0 \quad (21b)$$

$$-D \frac{\partial C}{\partial x} + VC \Big|_{x=0} = VC_c \quad 0 < t \leq t_o \quad (21c)$$

$$-D \frac{\partial C}{\partial x} + VC \Big|_{x=0} = 0 \quad t > t_o \quad (21d)$$

$$\left. \frac{\partial C}{\partial x} \right|_{x=L} = 0 \quad t > 0 \quad (21e)$$

The transport model and the associated initial and boundary conditions can be reduced to the same dimensionless equations derived for the two region model (Eqs. 6, 7, and 15). The dimensionless model is obtained using the following definitions for dimensionless variables

$$C_1 = \frac{C}{C_o} \quad C_2 = \frac{S_2}{(1 - F) K_d C_o} \quad (22)$$

$$T = \frac{Vt}{L} \quad (23)$$

$$P = \frac{VL}{D} \quad (24)$$

$$R = 1 + \frac{\rho K_d}{\theta} \quad (25)$$

$$\beta = \frac{\theta + F\rho K_d}{\theta + \rho K_d} \quad (26)$$

$$\omega = \frac{k_2 (1 - \beta) RL}{V} \quad (27)$$

3.1.3 Model 3. The One Site Model

In this nonequilibrium model, sorption is assumed to be completely rate limited, and it is approximated with first-order kinetics; thus, the governing equations to the transport model are

$$\theta \frac{\partial C}{\partial t} + \rho \frac{\partial S}{\partial t} = D\theta \frac{\partial^2 C}{\partial x^2} - v\theta \frac{\partial C}{\partial x} \quad (28)$$

and

$$\frac{\partial S}{\partial t} = k_3 (K_d C - S) \quad (29)$$

in which S is the sorbed species concentration, and k_3 is the desorption rate coefficient for a linear reversible sorption reaction. To approximate conditions of the miscible displacement experiments, Eqs. 26 and 27 should be solved under the following initial and boundary conditions

$$C(x, 0) = 0 \quad (30a)$$

$$S(x, 0) = 0 \quad (30b)$$

$$-D \frac{\partial C}{\partial x} + VC \Big|_{x=0} = VC_o \quad 0 < t \leq t_o \quad (30c)$$

$$-D \frac{\partial C}{\partial x} + VC \Big|_{x=0} = 0 \quad t > t_o \quad (30d)$$

$$\frac{\partial C}{\partial x} \Big|_{x=L} = 0 \quad t > 0 \quad (30e)$$

The one-site model can be expressed in the same convenient dimensionless form used previously with models 1 and 2. The dimensionless variables X , T , P , and R are defined as in model 2; however, the following dimensionless variables are also needed

$$C_1 = \frac{C}{C_o} \quad C_2 = \frac{S}{K_d C_o} \quad (31)$$

$$\beta = \frac{\theta}{\theta + \rho K_d} = \frac{1}{R} \quad (32)$$

$$\omega = \frac{k_3 L (1 - \beta) R}{V} \quad (33)$$

Likewise, with appropriate variable substitutions, the initial and boundary conditions transform to the same dimensionless form obtained with models 1 and 2. By expressing model 3 in the same dimensionless form as model 2, it becomes evident that the two models approach equivalency as β approaches the reciprocal of the retardation factor. β equals the reciprocal of R when, F , the mass fraction of instantaneous sorption sites is zero. It is also possible for β to approach a value of $1/R$ if F is less-than-or-equal to the reciprocal of R and the porous media has sufficiently high sorptive capacity.

3.2 MATERIALS and METHODS

3.2.1 Column Experiments

Breakthrough curves (BTCs) of pentafluorobenzoic acid (PFBA), benzene, xylene, PCE, toluene, and 1-methylnaphthalene were measured during aqueous elution through soil columns totally saturated with water and through soil columns containing water and a residual decane saturation. Fig. 2 shows a schematic of the experimental setup. The soil used was a fine sand. A sieve analysis on the sand determined the effective size and uniformity coefficient to be 0.24 mm and 1.67, respectively. The density of the sand grains was estimated at 2.68 g/cm^3 . The organic mass fraction of the sand was determined using a LECO 112 carbon determinator and found to be 0.02 percent.

Soil was packed in stainless steel columns with an inside diameter of 2.12 cm and with lengths ranging from 6.78 to 7.11 cm. All valves and tubing were stainless steel. The soil columns were saturated with a solution of 0.01 N CaCl_2 containing 10 mg/l mercuric chloride. Soil columns were initially exposed to several column pore volumes of the CaCl_2 solution until saturated. An aqueous solution of 100.0 mg/l PFBA and .01 N CaCl_2 was used to characterize pore water velocities, column void volumes, column Peclet numbers, and natural soil dispersivities. Next, a decane residual was emplaced on the soil matrix using a procedure of first eluting the columns with several pore volumes of decane, followed by several pore volumes of the CaCl_2 solution. Column elutions with the CaCl_2 solution continued until no residual decane was detected in the eluent.

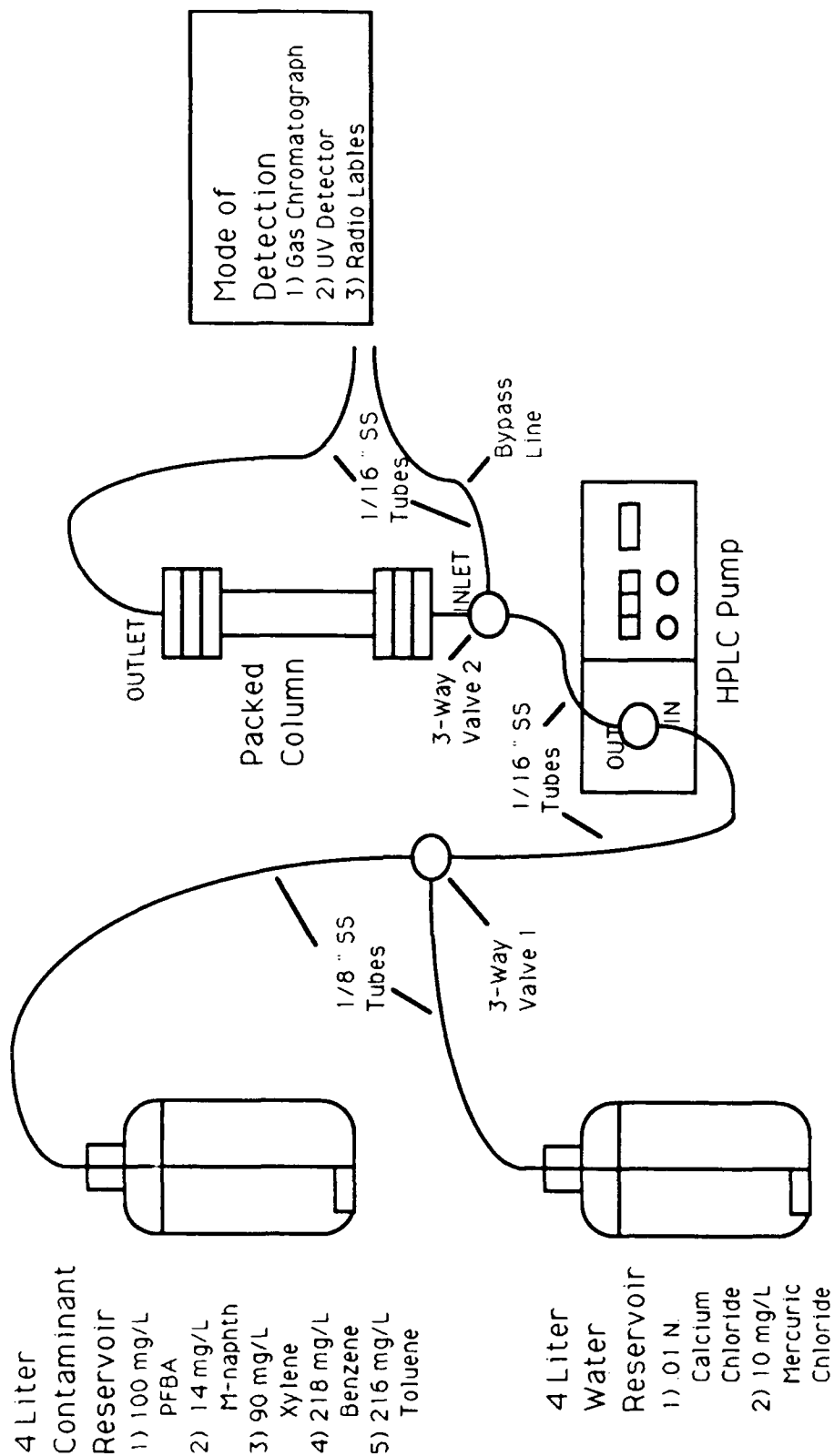


Fig. 2 Schematic of Experimental System.

Again, an aqueous solution of PFBA, was used to characterize the column void volumes, pore water velocities, Peclet numbers and dispersivities in the soil-SNAPL columns. Table 1 summarizes the soil column characteristics obtained by measurement and through modeling of PFBA BTCs. Benzene, toluene, PCE, xylene, and 1-methylnaphthalene were individually eluted as dissolved solutes in the CaCl_2 solution. The organic solution concentrations were generally less than 50 percent of reported aqueous-phase solubilities.

3.2.2 Saturation Flask Experiments and Residual Decane Determinations

Decane-water partition coefficients were determined for each reactive tracer used in the study. Following the approach of MacIntyre and Stauffer (1988), saturation flasks containing 350 ml of distilled water were prepared for each organic tracer. Next, 25 ml of a solution, containing 49.5 ml decane and 0.5 ml of tracer (i.e., benzene, toluene, PCE, xylene, or 1-methylnaphthalene), was carefully added to each flask above the water surface. Each flask was allowed to sit undisturbed for one week, after which tracer concentrations were analyzed in both water and decane phases using GC. Sampling and subsequent analyses were repeated three times for each tracer. The decane-water partition coefficients were calculated using the following equation

$$K_p = \frac{[(\text{mass of tracer in decane})/(\text{total mass of decane})]}{[(\text{mass of tracer in water})/(\text{total mass of water})]} \quad (34)$$

All soil columns were sacrificed and analyzed for decane after the completion of planned miscible displacement experiments. The stainless steel column end fittings were removed and the sand-decane-water contents were carefully sectioned into five equal samples. The decane was then extracted from the samples with hexane. The hexane extracts were analyzed by GC to obtain the mass of decane extracted per unit mass of sand.

3.2.3 Model Parameter Estimation by Nonlinear Least-squares Inversion

Transport model parameters V , V_m , D , D_m , R , T_o , β , and ω were estimated using software developed by Parker and van Genuchten (1984). The

Table 1. Basic Soil Column Characteristics Obtained Through PFBA BTCs

Exp No.	Column No.	Decane Present	Dry Bulk Density g/cm ³	Flux q cm/min	Method 1*		Method 2**	
					Water Content θ	Decane Content θ_0	Water Content θ	Decane Content θ_0
1	1	N	1.70709	.84916	.36303	NA	.36643	NA
2	1	N	1.70709	.28305	.36303	NA	.36253	NA
3	1	N	1.70709	.14157	.36303	NA	.36056	NA
4	2	N	1.72000	.28305	.35821	NA	.35815	NA
5	2	N	1.72000	.14157	.35821	NA	.35789	NA
6	2	Y	1.72000	.84916	.28856	.069649	.29225	--
7	2	Y	1.72000	.28305	.28856	.069649	.29476	.063387
8	2	Y	1.72000	.14157	.28856	.069649	.29243	.065462
9	3	N	1.66667	.28305	.37814	NA	.37724	NA
10	3	Y	1.66667	.26963	.33559	.042542	.31487	.062376
11	4	N	1.64043	.26710	.38790	NA	.38790	NA
12	4	Y	1.64043	.27773	.35112	.036783	.31976	.068142
13	5	N	1.6205	.28305	.39534	NA	.39534	NA
14	5	Y	1.6205	.28305	.33695	.05836	.33695	.05836
15	6	Y	1.7109	.84916	.29751	.06424	.29751	.06424
16	6	Y	1.7109	.28305	.29751	.06424	.29751	.06424
17	6	Y	1.7109	.14157	.29751	.06424	.29751	.06424

N = No, Y = Yes, NA = Not Applicable

* Calculated from the measured density of sand and the measured residual decane

** Calculated from results of fitting model 1 to PFBA BTCs

software uses nonlinear least-squares inversion to identify parameter values that minimize the sum deviations between analytical model estimates (i.e., model 1, 2 or 3) and observed solute concentrations. The software was originally developed to fit models one and two; however, parameter values for model 3 can be identified using this software if C_1 , C_2 , β , and ω are defined as specified in Eqs. 31 through 33.

PFBA has been used successfully in the past as a nonreactive tracer; thus, asymmetry or tailing in PFBA BTC would indicate the presence of transport related nonequilibrium conditions. Model 1 was used to simulate BTCs for PFBA, which in turn, identified characteristic pore water velocities, dispersion coefficients, and Peclet numbers for each soil column. In subsequent efforts to identify model 2 parameters for reactive tracers, column Peclet numbers, from corresponding PFBA experiments, were used in the model fitting process but not allowed to vary; this left parameters R , T_0 , β , and ω to vary.

Initial estimates of parameters are needed to begin the iterative search for values giving the best model fit. First estimates were obtained using solute concentration moments from BTCs. Using the same notation as Valocchi (1985), the n th absolute temporal moment for the BTC is defined as

$$m_n = \int_0^{\infty} T^n C_1(X, T) dT \quad (35)$$

in which T and X are the dimensionless time and distance variables defined previously. Having estimates of the soil column porosity and the rate at which water is eluted through the column, the zero absolute moment gives the total mass of contaminant injected per unit of column void volume. When working with dimensionless concentrations, the zero moment gives the length of the injection pulse, T_0 , expressed in column pore volumes. The n th normalized absolute moment is expressed as

$$\mu'_n = \frac{m_n}{m_0} = \frac{\int_0^{\infty} T^n C_1(X, T) dT}{\int_0^{\infty} C_1(X, T) dT} \quad (36)$$

The first normalized moment of the BTC represent the time on the BTC where half the solute mass has been eluted from the column. The retardation factor, R , was obtained from the first normalized moment. The n th central moment is defined as

$$\mu_n = \frac{\int_0^{\infty} (T - \mu_1)^n C_1(X, T) dT}{\int_0^{\infty} C_1(X, T) dT} \quad (37)$$

The second and third central moments were used to obtain estimates of β and ω . Valocchi (1985) gives moment equations for dirac contaminant loads. Used in this study, but derived from Valocchi's equations, are moment equations appropriate for pulse contaminant injections in one dimensional miscible displacement experiments (see Table 2).

3.3 RESULTS and DISCUSSION

3.3.1 Nonreactive Tracer Displacement

PFBA was used as a nonreactive tracer to characterize the physical transport parameters in soil columns both before and after residual decane emplacement. Table 3 summarizes the optimum parameter values obtained from fitting model 1 to the BTCs of PFBA. For several soil columns, miscible displacement experiments were conducted over a range of fluxes in order to observe model parameter dependency on pore water velocities. Fig. 3 shows typical BTCs for PFBA from a column before and after an emplaced decane residual. This figure shows the accelerated arrival of the tracer after a decane residual was created in the soil column. Under the same fluid flux, the presence of residual decane increased pore water velocities and dispersivities (α_{cs}); however, column Peclet numbers consistently decreased after decane emplacement. Decane saturation, θ_d , did not appear to have any consistent effect on the fraction of mobile water. Values of β were less than one but generally greater than 0.90.

An analysis was performed to evaluate the sensitivity of the general dimensionless model under various values of β and ω . The analysis lead to the derivation of the following equation

Table 2. Time Moment Formulas for Pulse Input

Moment	Model
	1, 2 and 3
μ'_1	$XR + \frac{T_o}{2}$
μ_2	$\frac{2XR^2}{P} + \frac{2X(1 - \beta)^2 R^2}{\omega} + \frac{T_o^2}{12}$
μ_3	$\frac{12XR^3}{P^2} + \frac{12X(1 - \beta)^2 R^3}{P\omega} + \frac{6X(1 - \beta)^3 R^3}{\omega^2}$

Table 3. Optimum Parameter Values for Model 1 Fitted to Observed BICs of PFBA

Exp No.	Column No.	Decane Present	Flux q cm/min	V cm/min	α_0 cm	P	β	ω
1	1	N	.84916	2.31736	.052991	127.945	.98446	.08759
2	1	N	.28305	.78078	.052384	129.43	.98450	.01015
3	1	N	.14157	.39252	.052227	129.82	1.0	0.0
4	2	N	.28305	.79033	.036149	187.56	.96132	.21499
5	2	N	.14157	.39545	.036945	183.52	.96206	.24132
6	2	Y	.84916	2.90562	.038477	176.21	.92752	.52716
7	2	Y	.28305	.96028	.038843	174.55	.93723	.28458
8	2	Y	.14157	.48397	.038432	176.42	.93621	.2409
9	3	N	.28305	.75032	.037717	179.76	.98702	.70035
10	3	Y	.26963	.85633	.071199	95.23	.99582	.19197
11	4	N	.26710	.68935	.015812	448.18	.97000	.48917
12	4	Y	.27773	.86860	.055964	126.63	.97041	.09600
13	5	N	.28305	.7160	.059134	118.12	.9900	.24498
14	5	Y	.28305	.84003	.068188	102.44	.91829	.01053
15	6	Y	.84916	2.854	.09713	71.93	.97962	.05464
16	6	Y	.28305	.9514	.09712	71.93	.99142	.00956
17	6	Y	.14157	.4757	.09712	71.93	.98867	.00466

N = No, Y = Yes

BTC EXPERIMENTS 9 AND 10

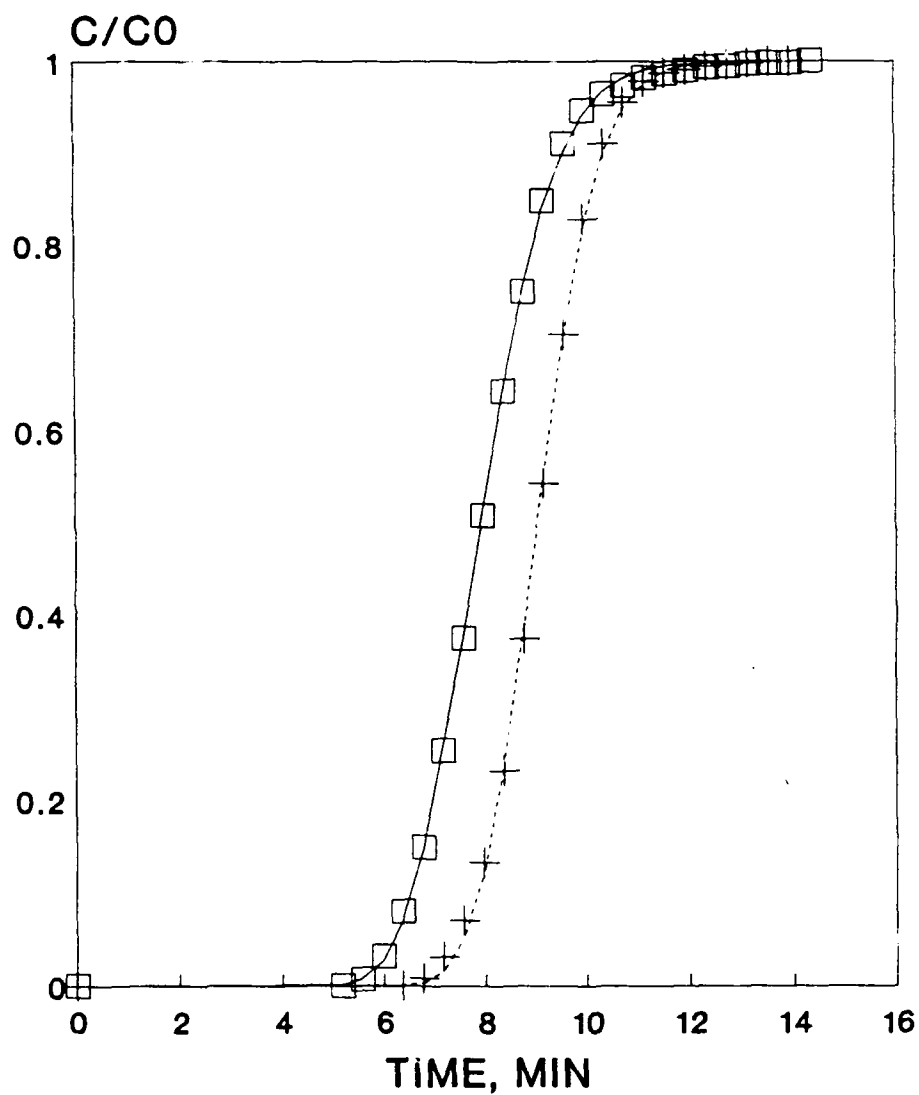


Fig. 3 PFBA BTCs before (+ observed, --- fitted) and after (□ observed, — fitted) Emplacement of Decane Residual on Soil Column No. 3.

$$-\frac{\Delta\beta}{\beta} \left(\frac{\beta}{1-\beta} \right) = \frac{\Delta\omega}{\omega} \quad (38)$$

which describes the sensitivity of the general dimensionless model to small changes in the relative values of β and ω . According to Eq. 38, as β values approach unity the transport model becomes less responsive to changes in the value of ω . This conclusion was verified through simulation. Fig 4 shows how the model responds to relative changes in ω , if the standard ω is 2.0. As the true β approaches unity, the response curves become flatter. The flat response curves could make some ω determinations less reliable since a model fitting procedure was used. Table 3 lists several fitted values of ω that may be suspect because β values were close to 1.0.

3.3.2 Reactive Tracer Displacement

Table 4 summarizes miscible displacement experiments conducted with reactive solutes, the columns used, whether residual decane was present, the mode of tracer detection, the final models used to simulate BTCs, and the fitted model parameters. Looking first, at reactive tracer behavior in the natural soil (experiments 18 through 32), the retardation factors were generally less than 1.5 with the exception of 1-methylnaphthalene. For benzene and toluene, β values were greater than .90. Values of β decreased with increasing tracer hydrophobicity, while fitted values of ω consistently increased (see Table 4, experiments 18 through 32). Fig. 5 shows typical BTCs for benzene and toluene from a column before treatment with decane. Given the small fraction of immobile water found from the PFBA BTCs, and the increasing asymmetry seen in the BTCs of less polar tracers, it was assumed that the observed nonequilibrium was sorption related. Thus, model 2 was selected to approximate reactive solute transport through the water saturated soil columns. Changes in the fluid velocity did not produce consistent variations in ω values. The value of F was estimated by regressing the product $\theta \cdot R \cdot \beta$ against $\theta \cdot (R-1)$ (see Fig. 6). The regression equation indicates that 33 percent of the sorption sites were associated with instantaneous sorption. Another correlation

DIMENSIONLESS MODEL SENSITIVITY TO OMEGA

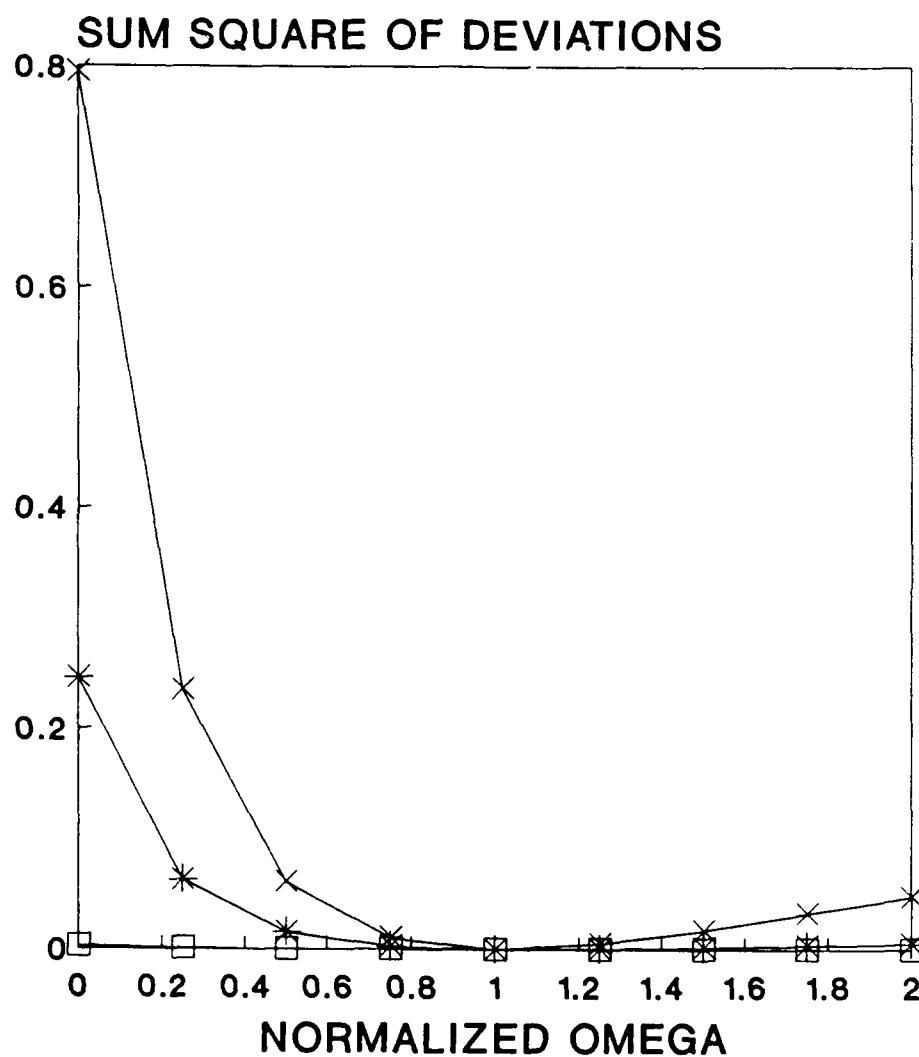


Fig. 4 Dimensionless Model Sensitivity to ω under Specified Values of β
(□, $\beta = .99$; *, $\beta = .90$; and x, $\beta = .70$).

Table 4. Optimum Parameter Values for Models 1 and 3 Fitted to Observed BTCs for Reactive Organic Tracers

Exp No.	Column No.	Decane Present	Chemical	Detection Mode	Model No.	Flux q cm/min	Fitted Values from Model			
							P	R	β	ω
18	1	N	B	UV	2	.84916	127.95	1.04427	.96241	.07380
19	1	N	B	UV	2	.28305	129.43	1.02398	.96884	.05352
20	1	N	B	UV	2	.1415	129.82	1.02049	.98873	.02016
21	1	N	T	UV	2	.84916	127.945	1.10517	.96070	.34164
22	1	N	T	UV	2	.28305	129.43	1.09147	.96735	.15473
23	1	N	T	UV	2	.14157	129.82	1.13821	.94140	.06508
24	1	N	X	UV	2	.84916	127.95	1.39889	.80035	1.40666
25	1	N	X	UV	2	.28305	129.43	1.44777	.82351	1.31820
26	1	N	X	UV	2	.14157	129.82	1.44548	.83125	1.56486
27	1	N	P	UV	2	.84916	127.95	1.47472	.77739	1.36001
28	1	N	P	UV	2	.28305	129.43	1.49704	.79466	1.55929
29	1	N	P	UV	2	.14157	129.82	1.50594	.80914	1.80547
30	1	N	M	UV	2	.84916	127.95	3.36166	.49267	1.69485
31	1	N	M	UV	2	.28305	129.43	3.40073	.53596	2.20321
32	1	N	M	UV	2	.14157	129.82	3.61973	.5455	2.29345
33	2	Y	B	UV	3	.84916	176.21	40.59381	.02460	4.69985
34	2	Y	B	UV	3	.28305	174.55	40.44733	.02470	6.48095
35	2	Y	B	UV	3	.14157	176.42	40.42116	.02470	9.96527
36	2	Y	T	UV	3	.84916	176.21	177.0383	.00560	4.59269
37	2	Y	T	UV	3	.28305	174.55	179.46825	.00560	6.89996
38	2	Y	T	UV	3	.14157	176.42	169.15788	.00590	10.03995
39	3	Y	X	UV	3	.28305	95.23	624.70735	.00160	4.60118
40	4	Y	P	SC	3	.28305	126.63	725.85339	.00138	7.88639
41	5	Y	M	GC	3	.28305	102.44	2145.88112	.00047	3.89892
42	6	Y	M	GC	3	.116192	71.92	2394.2773	.00042	6.28484

N = No, Y = Yes, UV = Ultra Violet, SC = Scintillation Counter, GC = Gas Chromatography,

B = Benzene, T = Toluene, X = Xylene, P = PCE, M = 1-methylnaphthalene

BTC EXPERIMENTS 18 AND 21

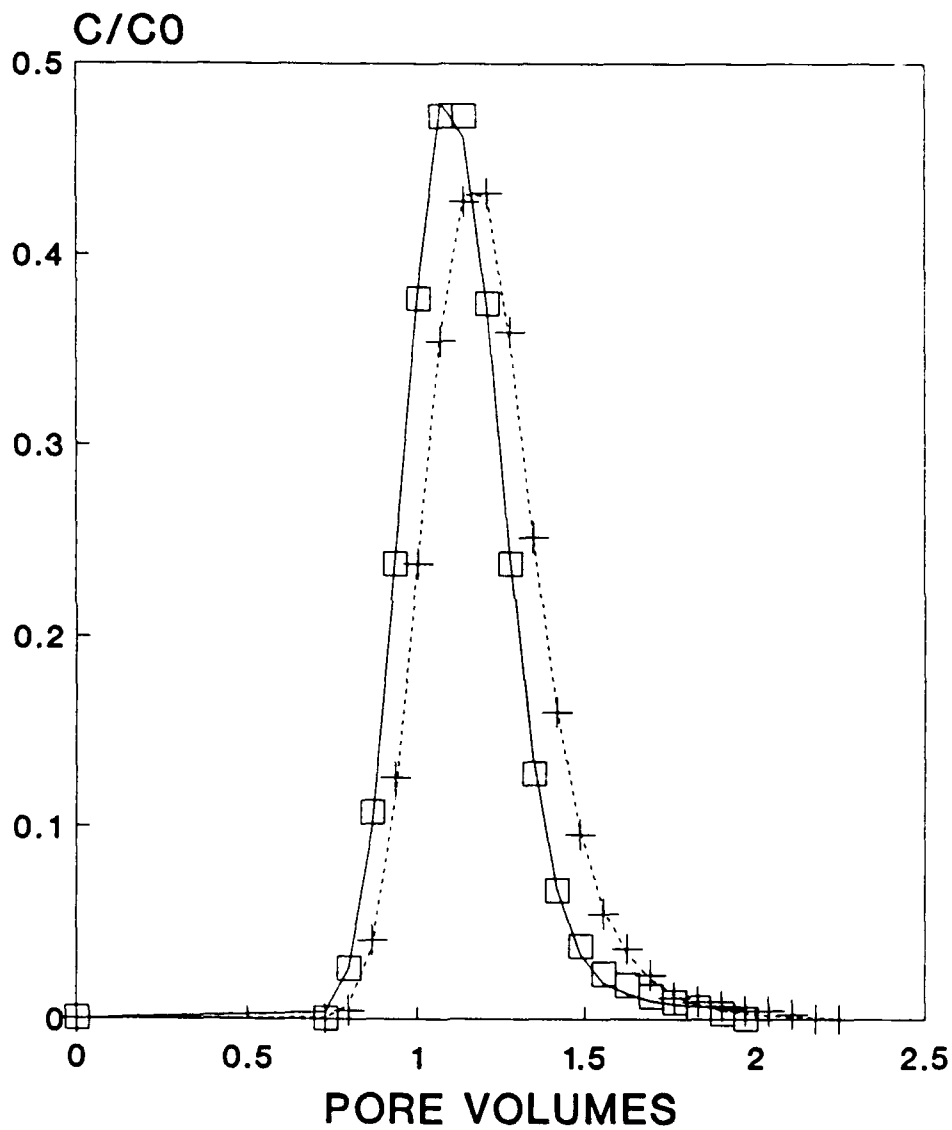


Fig. 5 BTCs of Benzene (\square observed, — fitted) and Toluene (+ observed, --- fitted) before Emplacement of Decane Residual on Soil Column No. 1.

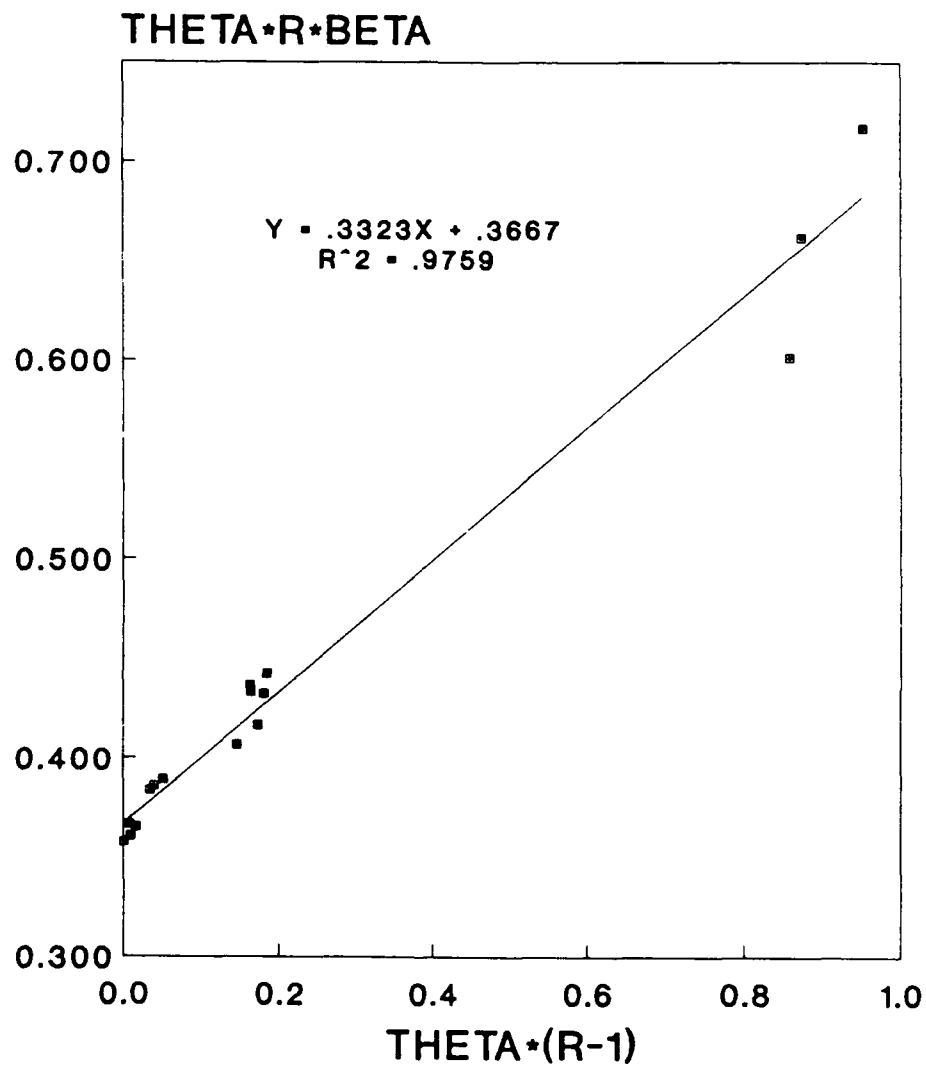


Fig. 6 Regression of $\theta R\beta$ against $\theta(R-1)$ for all Reactive and Nonreactive Tracers before Emplacement of Decane Residual on Soil Column No. 1.

was sought between logarithmic values of the first order rate coefficient k_2 and the sorption coefficient, K_d . Log linear relationships were visible for xylene, PCE, and 1-methylnaphthalene where experiments were conducted at the same flux (see Fig. 7). Poor correlations were obtained for toluene and benzene regardless of the flux. In light of the β - ω relationship defined by Eq. 38, the high β values probably precluded suitable identification of ω in experiments 18 through 23.

The emplacement of residual decane on the soil columns increased tracer retardation factors 40 to 650 times. Fig. 8 shows BTCs of benzene and toluene after treating a soil column with decane; this was the same experimental column used previously to produce the BTCs depicted in Fig. 5 for decane free soil.

Initial efforts to use model 2 to predict BTCs were not successful. In general, the fitting program forced β to zero, suggesting instantaneous sorption was a minor component of the total sorptive capacity of SNAPL-Soil matrix. This indicates that solute exchange between water and decane was for the most part rate limiting, and that model 3 could be applicable. In all subsequent model fitting efforts, model 3 was successfully used to predict BTCs for reactive tracers from soil columns containing residual decane.

Another correlation was sought, but this time between the logarithmic values of the first order rate coefficient, k_3 , and the sorption coefficient, K_d . Log linear relationships were again visible between tracer experiments conducted at the same flux (see Fig. 9); however, an overall correlation was taken. Limited data precludes development of separate regression equations for each fluid flux; nevertheless, future efforts to develop such correlations should consider an array of experiments using a range of fluid fluxes and a host of tracers with a broad range of K_d s.

Explanations were sought for why model 3 worked on the soils containing residual decane. One reason, could be that most of the sorption capacity was located in randomly distributed but isolated droplets of SNAPL. If this were true, it could be argued for low organic soils, that the fraction of instantaneous sorption sites was essentially zero. When

k₂ VS. K_d

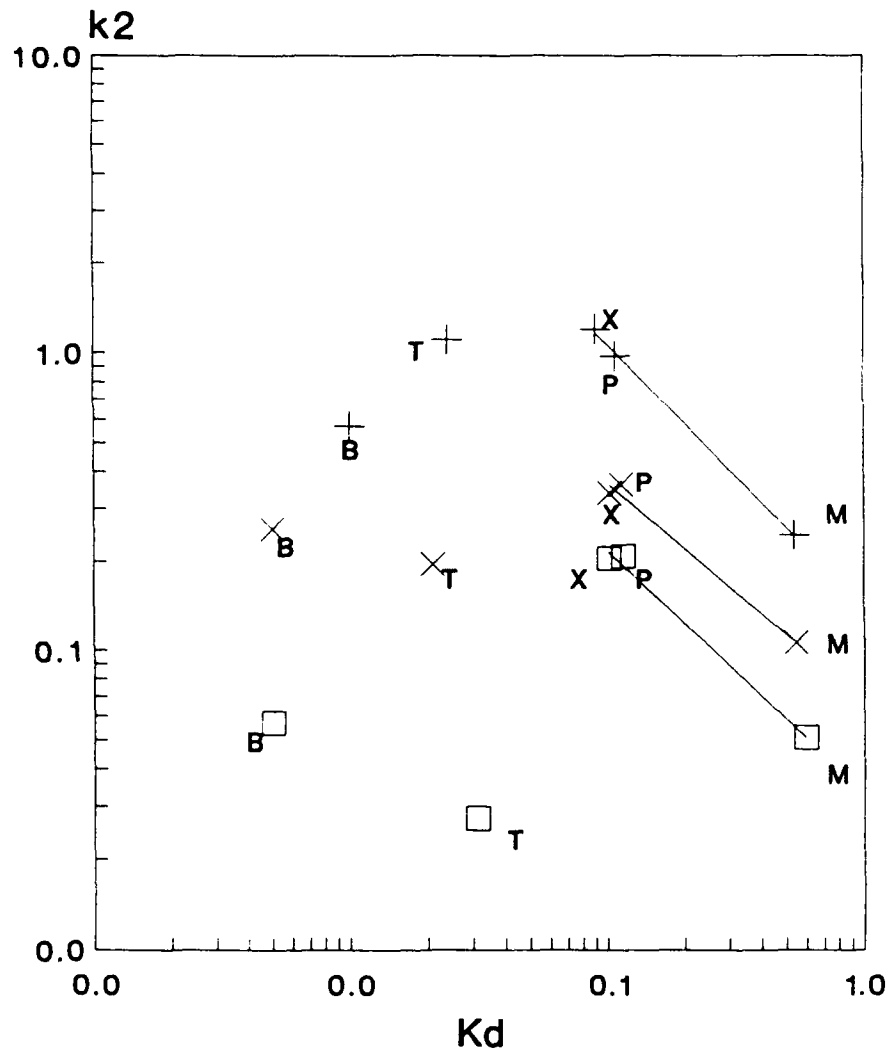


Fig. 7 First Order Rate Coefficient, k_2 versus Soil Sorption Coefficient, K_d before Emplacement of Decane Residual on Soil Column No. 1 (B = benzene, T = toluene, X = xylene, P = PCE, and M = 1-methylnaphthalene; at fluid fluxes, \square = .1415 cm min⁻¹, \times = .283/cm min⁻¹, and $+$ = .8492 cm min⁻¹).

BTC EXPERIMENTS 33 AND 36

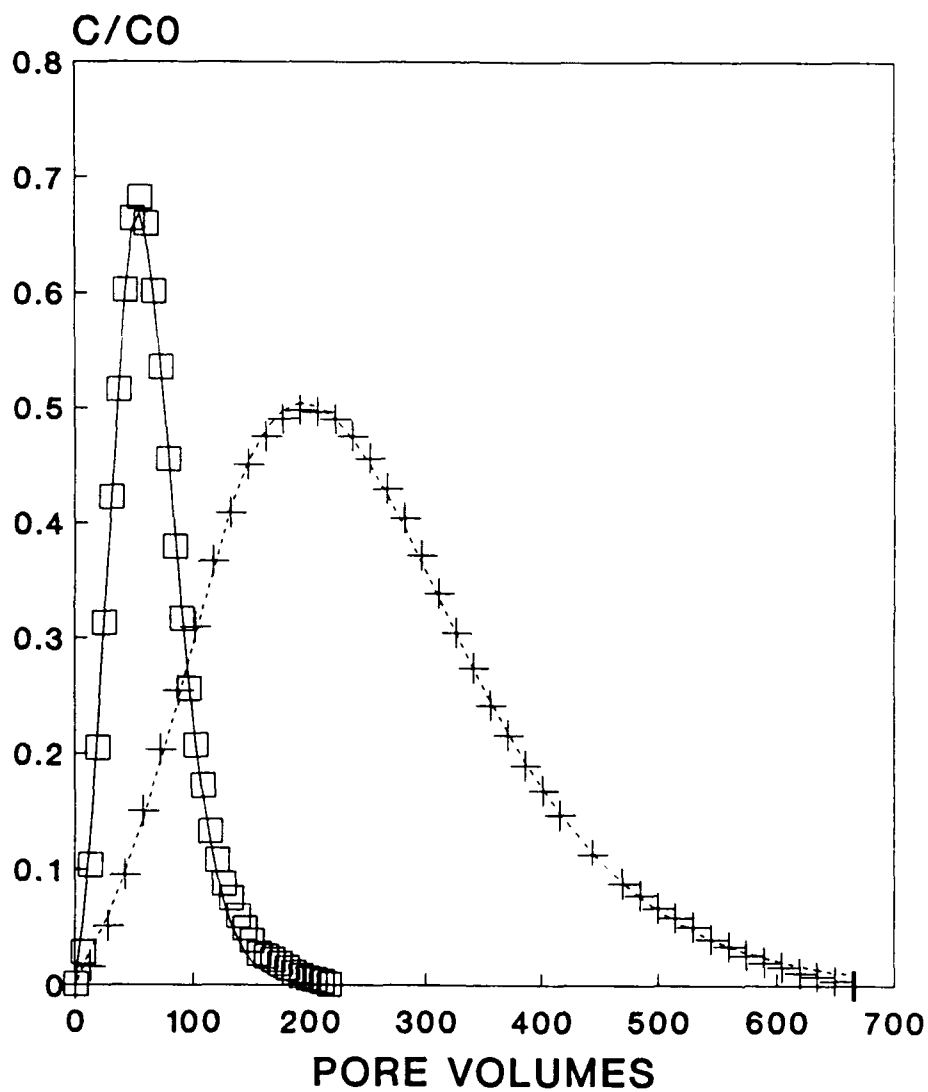


Fig. 8 BTCs of Benzene (\square observed, — fitted) and Toluene (+ observed, --- fitted) after Emplacement of Decane Residual on Soil Column No. 2.

k3 VS. Kd

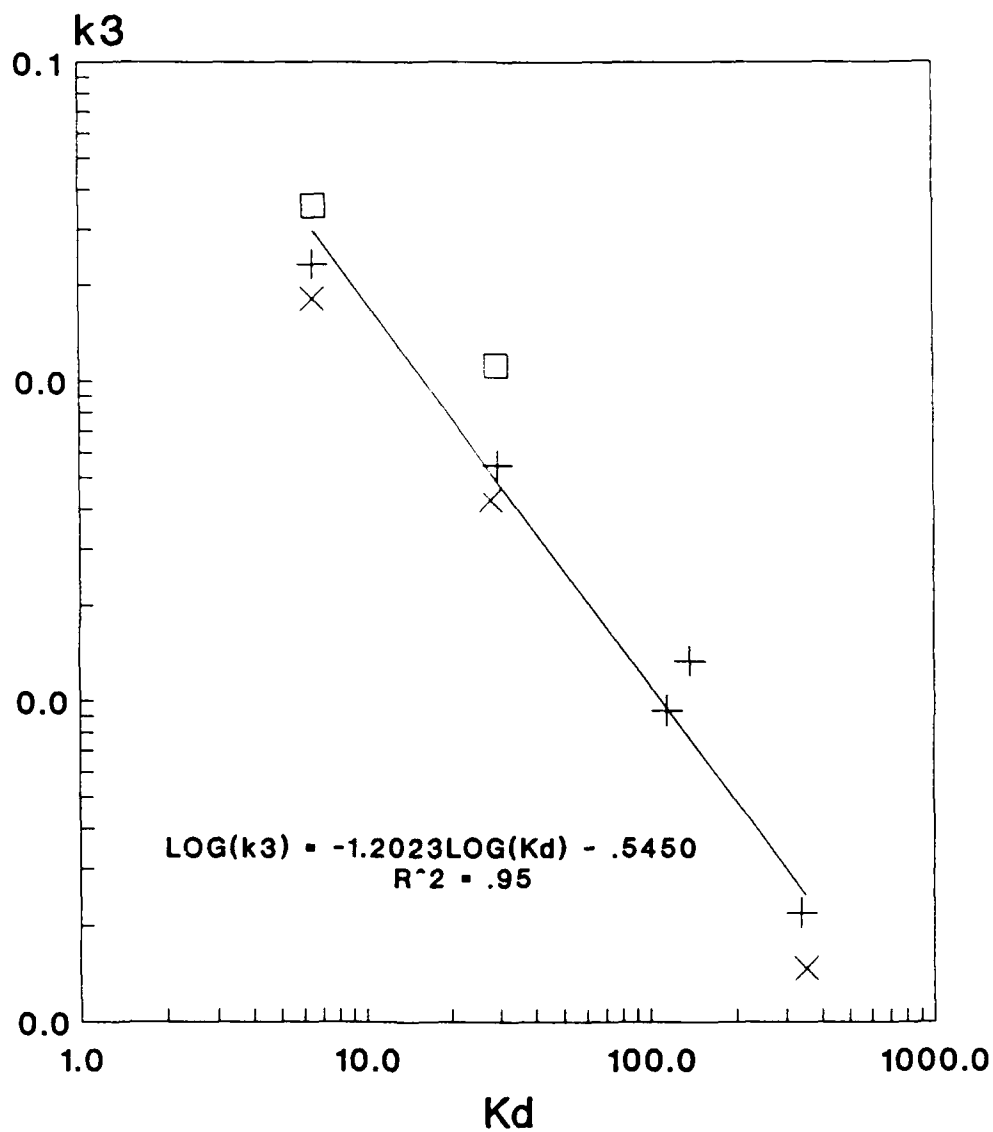


Fig. 9 First Order Rate Coefficient, k_3 versus SNAPL-Soil Sorption Coefficient, K_d after Emplacement of Decane Residual. Exp. Nos. 33 to 42 for Fluid Fluxes ($\times = .1416 \text{ cm min}^{-1}$, $+$ = .2831 cm min^{-1} , and $\square = .8492 \text{ cm min}^{-1}$).

F is zero, model 2 reduces to model 3 and β equals $1/R$. This explanation is plausible, if tracer exchange is restricted by the available area through which solutes pass between the aqueous and organic phases.

In the preparation of soil columns, decane was eluted through soils which had been previously saturated with water. The elution of decane entrapped water in the smaller pores. In a similar manner, when water was subsequently used to displace excess decane, a residual organic phase could have been entrapped as semi-isolated droplets between grains of sand.

Another possible explanation for the success of model 3, could be that F is not zero; rather, the sorption capacity of the SNAPL-soil matrix is so large, that β (from model 2) approaches the value of F . If the value of F is less-than-or-equal-to $1/R$, then predictions from model 2 would either equal or closely resemble predictions from model 3.

3.4 INDEPENDENT MODEL PARAMETER ESTIMATION

It is often the case with modeling that parameters values are needed which cannot be measured conveniently or directly; consequently, it is desirable to develop a capability of estimating model parameters independently of experiment. In this study, β , ω , P , and R were the most important model parameters for reactive tracers. In experiments with soil columns partially saturated with decane, sorption had such a dominant influence on the shape of BTC, that it masked the influence of the Peclet number.

In this study an effort was made to estimate β , ω , and R independently; however, the greatest success was with the estimation of R and β . Table 5 shows the values of R estimated by two independent methods, and my moments analysis. Both of the independent methods used measured decane-water partition coefficients and sorption coefficients for the sand. The methods differ in the manner by which the available void volume and the residual decane is estimated. In method 1, the residual decane and the column void volume are determined from the estimated pore velocities given by the PFBA BTCs. Method 2 makes use of the column void volumes and decane residuals obtained through measured soil densities and

Table 5. Optimum Fitted Retardation Factors versus Independent Estimates and Calculated Retardation Factors from Moments

Retardation Factor R					
Exp No.	Chemical	Flux q cm/min	Independent Estimates		Model Fit
			Method 1**	Method 2***	
18	B	.84916	NA	NA	1.0420
19	B	.28305	NA	NA	1.0169
20	B	.1415	NA	NA	1.02083
21	T	.84916	NA	NA	1.1214
22	T	.28305	NA	NA	1.0937
23	T	.14157	NA	NA	1.0899
24	X	.84916	NA	NA	1.4394
25	X	.28305	NA	NA	1.4667
26	X	.14157	NA	NA	1.4562
27	P	.84916	NA	NA	1.5064
28	P	.28305	NA	NA	1.5066
29	P	.14157	NA	NA	1.5085
30	M	.84916	NA	NA	3.5842
31	M	.28305	NA	NA	3.5381
32	M	.14157	NA	NA	3.7684
33	B	.84916	43.1835	47.3248	42.37978
34	B	.28305	43.1835	47.3248	42.16637
35	B	.14157	43.1835	47.3248	40.35581
36	T	.84916	163.7353	179.7175	172.7525
37	T	.28305	163.7353	179.7175	175.19283
38	T	.14157	163.7353	179.7175	164.0406
39	X	.28305	632.3564	405.1020	579.3803*
40	P	.84916	725.8185	357.5349	NA
41	M	.28305	1669.2600	1669.2600	NA
42	M	.11619	2080.7126	2080.7126	2394.2773

B = Benzene, T = Toluene, P = PCE, X = Xylene, M = 1-methylnaphthalene, NA = Not Applicable

* Estimated from the integral $\int [1 - \frac{C}{C_0}] dt$

** Using PFBA BTCs, K_{ps} from saturation flasks, and natural soil K_d s

*** Using measured soil densities and measured residual decane

direct measurement of residual decane. Once the decane residual and the column void volume have been estimated, the calculation of R is the same between the two methods. When a comparison is drawn between independently derived R s and those obtained through model fit, it may be concluded that both independent methods are good estimators. Moment estimates were also excellent; however, this method requires an experimental effort that independent methods avoid.

The successful estimation of retardation factors meant that good estimates of β were likely; this was because the BTCs from experiments 33 to 42 were easily modeled as one site systems. Recall that in model 3, β is estimated simply from the reciprocal of the retardation factor. Table 6 shows independently estimated β values for model 3 as well as those obtained through model fitting; again a comparison will verify that methods of independent determination are quite reliable.

When β values are as small as those reported in table 6, it is evident from Eq. 38, that any model response to a change in this parameter is proportional to the value of β . Given in table 6 are β values calculated as if model 2 was being used. It should be noted that β values calculated for model 2 are not that different from the β estimates for model 3; this is because, large relative changes in this parameter will have little influence on simulated results if the true value of β is small.

4.0 PREDICTING SORPTION SYSTEM PERFORMANCE THROUGH EFFECTIVE RETARDATION

4.1 THEORY

Contaminant fluxes can be estimated from pollutant concentrations and the velocity of the groundwater flow. Contaminant flux reductions achieved by an SSS depend on the spatial distribution of SNAPL which controls spatial variations in the sorptive capacity and permeability of the soil. Designing an SSS for the field is complicated because several parameters affecting system performance require simultaneous consideration. The most salient parameters are: the residual SNAPL saturation, the permeability of the system and the surrounding aquifer, the geometrical configuration of the SSS, and the sorption characteristics of the natural aquifer and the sand/SNAPL matrix. The complexity of the problem is

Table 6. Optimum β Values for Model 3 Versus Independent Estimates for Models 2 and 3

Independent Estimates of β							Fitted β for Model 3
Exp No.	Chemical	For Model 3		For Model 2			
		Method 1*	Method 2**	Method 1*	Method 2**		
33	B	.023157	.021131	.0010495	.0009556	.02460	
34	B	.023157	.021131	.0005685	.0004423	.02470	
35	B	.023157	.021131	.0004857	.0004423	.02470	
36	T	.006107	.005564	.0006463	.0005951	.00560	
37	T	.006107	.005564	.0005621	.0005118	.00560	
38	T	.006107	.005564	.0008493	.0007733	.00590	
39	X	.001581	.002469	.0007092	.0010806	.00160	
40	P	.001378	.002797	.00068574	.00139409	.00138	
41	M	.000599	.000599	.0014750	.0014750	.00047	
42	M	.000481	.000481	.00118320	.00118320	.00042	

B = benzene, T = Toluene, X = Xylene, P = PCE, M = 1-methylnaphthalene

* Using PFBA BTCs, measured K_p s from saturation flasks, and natural soil K_d s

** Using measured sand densities, decane residuals K_p s from saturation flasks, and natural soil K_d s

reduced, however, if remediation benefits obtained from an SSS are expressed in terms of the 'effective retardation'. The effective retardation is defined here as the dimensionless ratio of contaminant velocity under ambient conditions, to the new velocity created with the installation of the SSS. The effective retardation is a spatially varying parameter of the groundwater flow domain; it is the factor that translates ambient contaminant velocities to what they will be once the sorption system is installed.

General analytical expressions for the effective retardation at all points within a flow domain are derived for the hypothetical sorption system illustrated in Fig. 10; this figure shows a circular sorption field centered over the origin of an orthogonal pair of axes. If a profile view were given, it would verify that the system is actually a vertical cylinder having a height equal to the thickness of the saturated zone and a horizontal radial dimension R_s . Inside the SSS, the porous media has homogeneous sorptive characteristics, porosity, and permeability. The aquifer surrounding the sorption system is also homogeneous; however, the sorption characteristics, porosity, and permeability of the porous media are different from that of the soil-SNAPL matrix in the SSS.

Derivation of the effective retardation functions for the cylindrical system involves elementary manipulations with complex numbers. For convenience the location of the SSS is centered over the origin of complex coordinate plane (see Fig. 10). The region inside the boundaries of the system will be referred to as the SSS, the flow domain surrounding SSS will be designated simply the surrounding aquifer (SA).

The derivation of the effective retardation equations begins with an assumption that transport is primarily advective (i.e., dispersive transport is insignificant). Next, an allowance is made that contaminant sorption to the soil and SNAPL is linear, reversible, and at equilibrium. With these two initial assumptions, the transport of a contaminant is given by the local apparent velocity

$$V_c = \frac{\bar{q}}{R \theta} \quad (39)$$

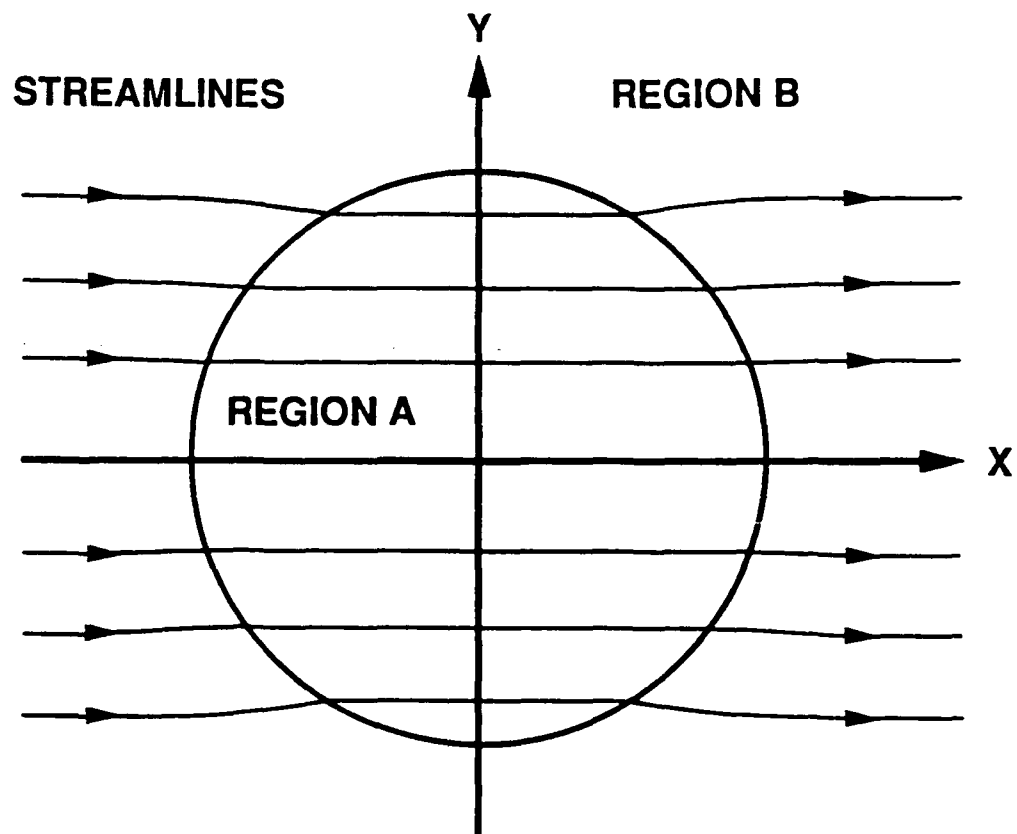


Fig. 10 Plan View of a Hypothetical Cylindrical SSS (Region A) Possessing Greater Permeability than the Surrounding Aquifer (Region B).

where, θ is the water saturated porosity and R is the retardation factor due to the sorption assumption. The retardation factor is the ratio of the average linear velocity of the groundwater to the velocity of a sorbing contaminant. \bar{q} is the resultant Darcian velocity obtained from discharge components in the horizontal x and y directions

$$\bar{q} = (q_x^2 + q_y^2)^{1/2} \quad (40)$$

The direction of this flow is ϕ

$$\phi = \text{Arctang} \left(\frac{q_y}{q_x} \right) \quad (41)$$

The retardation factor is estimated from

$$R = \left(1 + \frac{K_d \rho}{\theta} \right) \quad (42)$$

in which ρ is the bulk density of the permeable matrix, and K_d is the matrix/water sorption coefficient for a target contaminant. Though other formulations for the retardation factor exist (Hamaker 1975), Endfield (1982) produced a lucid derivation of Eq. 42 from assumptions that sorption was linear, reversible, and instantaneous.

Under natural soil conditions, the retardation factor is calculated by letting ρ equal ρ_s the bulk density of the soil, θ_s be the total soil porosity θ , and K_d equal K_{ds} the soil/water sorption coefficient for a target contaminant. The sorption coefficient may be obtained experimentally or estimated by any number of empirical formulae; however, a commonly used equation has been that from Karickhoff (1979)

$$K_{ds} = 0.63 K_{ow} F_{oc} \quad (43)$$

where K_{ow} is the octanol/water partition coefficient for the target contaminant and F_{oc} is the fraction of organic carbon associated with the soil.

Eq. 42 applies inside the SSS as well; however, values for θ and R must reflect the presence of the SNAPL. The water saturated porosity is obtained from knowing the SNAPL filled porosity θ_d .

$$\theta = \theta_s - \theta_d \quad (44)$$

The retardation factor inside the SSS is calculated from Eq. 42, using the newly determined water saturated porosity and values for ρ and K_d obtained from the following

$$\rho = \rho_s + \theta_d \rho_d \quad (45)$$

and

$$K_d = \frac{(K_{ds} \rho_s + K_p \theta_d \rho_d)}{(\rho_s + \theta_d \rho_d)} \quad (46)$$

in which ρ_d is the density of the SNAPL and K_p is the SNAPL/water partition coefficient for a targeted contaminant.

The effective retardation, R_e , at a location in the flow domain is the ratio of ambient contaminant velocity to the velocity after the SSS is installed. For locations inside the SSS and in the SA, respectively, the effective retardations are expressed as

$$R_e|_{SSS} = \frac{V_{co}}{V_c|_{SSS}} \quad (47a)$$

and

$$R_e|_{SA} = \frac{V_{co}}{V_c|_{SA}} \quad (47b)$$

where V_{co} is the ambient contaminant velocity, $V_c|_{SSS}$ is contaminant velocity in the SSS, and $V_c|_{SA}$ is the contaminant velocity in the SA.

In the development of the general effective retardation equations it is assumed that groundwater flow is uniform and parallel to the x axis in our hypothetical problem; hence, \bar{q} is initially equated to q_{x0} for the entire flow domain. It follows from this assumption, that the ambient contaminant velocity V_{f0} is found by combining Eqs. 39, 42, and 43.

$$V_{co} = \frac{q_{x0}}{\theta + 0.63 K_{ow} F_{oc} \rho_s} \quad (48)$$

After the sorption system is installed, the same substitutions described above define the apparent contaminant velocity in the SA as

$$V_{cl|SA} = \frac{\bar{q}}{\theta + 0.63 K_{ow} F_{oc} \rho_s} \quad (49)$$

Finally, the apparent contaminated velocity inside the SSS is obtained from combining Eqs. 39, 42, 44, 45 and 46.

$$V_{cl|SSS} = \frac{\bar{q}}{\theta - \theta_d + 0.63 K_{ow} F_{oc} \rho_s + K_p \theta_d \rho_d} \quad (50)$$

Analytical formulations for the effective retardation inside the SSS and in the SA can be obtained using appropriate expressions for the specific discharge. Strack and Haitjema (1981) provide a solution to a specific groundwater flow problem from which the required discharge expressions can be obtained. The solution is for steady flow in a regional two dimensional horizontal aquifer, that has a cylindrical homogeneous anomaly; the regional aquifer is characterized with a different hydraulic conductivity than the cylindrical anomaly. The solution incorporates far field boundary conditions of uniform flow parallel to the x axis and near field conditions of continuity of discharge along the boundary separating the SSS and the SA. For a unit thickness of aquifer, the complex potentials for groundwater flow in the SSS and the SA are

$$\Omega_{SSS} = -q_{x0} \frac{2 k_{SSS}}{k_{SA} + k_{SSS}} z + \frac{SSS}{\Phi_0} \quad (51a)$$

and

$$\Omega_{SA} = -q_{x0} \left[z + \frac{k_{SA} - k_{SSS}}{k_{SA} + k_{SSS}} \left(\frac{R_s^2}{z} \right) \right] + \frac{k_{SA}}{k_{SSS}} \frac{SSS}{\Phi_0} \quad (51b)$$

in which Ω_{SSS} and Ω_{SA} are the complex potentials inside the SSS and in the SA, respectively. The complex variable z is a point in the complex plane of which the origin of that plane is coincident with the center location of the SSS (see Fig. 10). The variable z equals $x + iy$, in which x and y are distance variables and i is the square root of negative one. k_{SSS} is the permeability of the SSS, k_{SA} is the natural permeability of the SA, and $\frac{SSS}{\Phi_0}$ is the product of initial potentiometric head and hydraulic conductivity inside the sorption system.

The derivative of the complex potential with respect to variable z can be used to obtain the horizontal specific discharge components. Inside the sorption system, the real portion of the derivative of Eq. 51a is

$$q_x = q_{x0} \left(\frac{2 k_{SSS}}{k_{SSS} + k_{SA}} \right) \quad (52a)$$

The imaginary portion is

$$q_y = 0 \quad (52b)$$

Eqs. 52a and 52b state that flow inside the boundary of the sorption zone is uniform and parallel to the x axis; thus,

$$\bar{q} = q_{x0} \left(\frac{2 k_{SSS}}{k_{SSS} + k_{SA}} \right) \quad (53)$$

The permeability of the sorption zone is a function of the residual saturation of SNAPL; therefore, k_{SSS} is the product of the relative permeability factor k_r , and the natural permeability of the soil k_h , used to construct the sorption field. Relative permeabilities have been estimated previously (Hatfield et al. 1991) for same soil used in the miscible displacement experiments. The function used then, to make estimates was from Rose (1949).

$$k_{rw} = \frac{16S_w^2 (S_w - S_{wm})^3 (1 - S_{wm})}{[2S_w^2 (2 - 3S_{wm}) + 3S_w S_{wm} (3S_{wm} - 2) + S_{wm} (4 - 5S_{wm})]^2} \quad (54)$$

in which S_{wm} is the residual saturation of water when decane/mineral oil is flowing through the media, and S_w is the ratio of water saturated porosity to total porosity. The residual saturation of water for the media under dynamic conditions was obtained from column experiments in which the SNAPL was used to displace water in a fully saturated system. Mass balance calculations gave an average residual water saturation of 0.246 for the columns.

By combining Eqs. 47a, 48, 50, and 53 and substituting the product $k_r k_h$ for k_{SSS} an analytical expression is obtained for the effective retardation inside the boundaries of the subsurface sorption system.

$$R_{e|SSS} = \frac{(\theta - \theta_d + 0.63 K_{ow} F_{oc} \rho_s + K_p \theta_d \rho_d) (k_r k_h + k_{SA})}{2 (\theta + 0.63 K_{ow} F_{oc} \rho_s) k_r k_h} \quad (55)$$

4.2 APPLICATION

Data from previous bench-scaled experiments (Hatfield 1989) were used to create hypothetical plots of the effective retardation inside the SSS (see Fig. 11). The plots cover a range of residual SNAPL saturations that exceed the maximum value of 13.9 percent observed experimentally (Hatfield 1989). It is expected that other SNAPLs are likely to produce effective retardations curves of similar shape, but may not be limited to the range of residual saturations found experimentally; thus, the curves

EFFECTIVE RETARDATION IN THE SORPTION SYSTEM

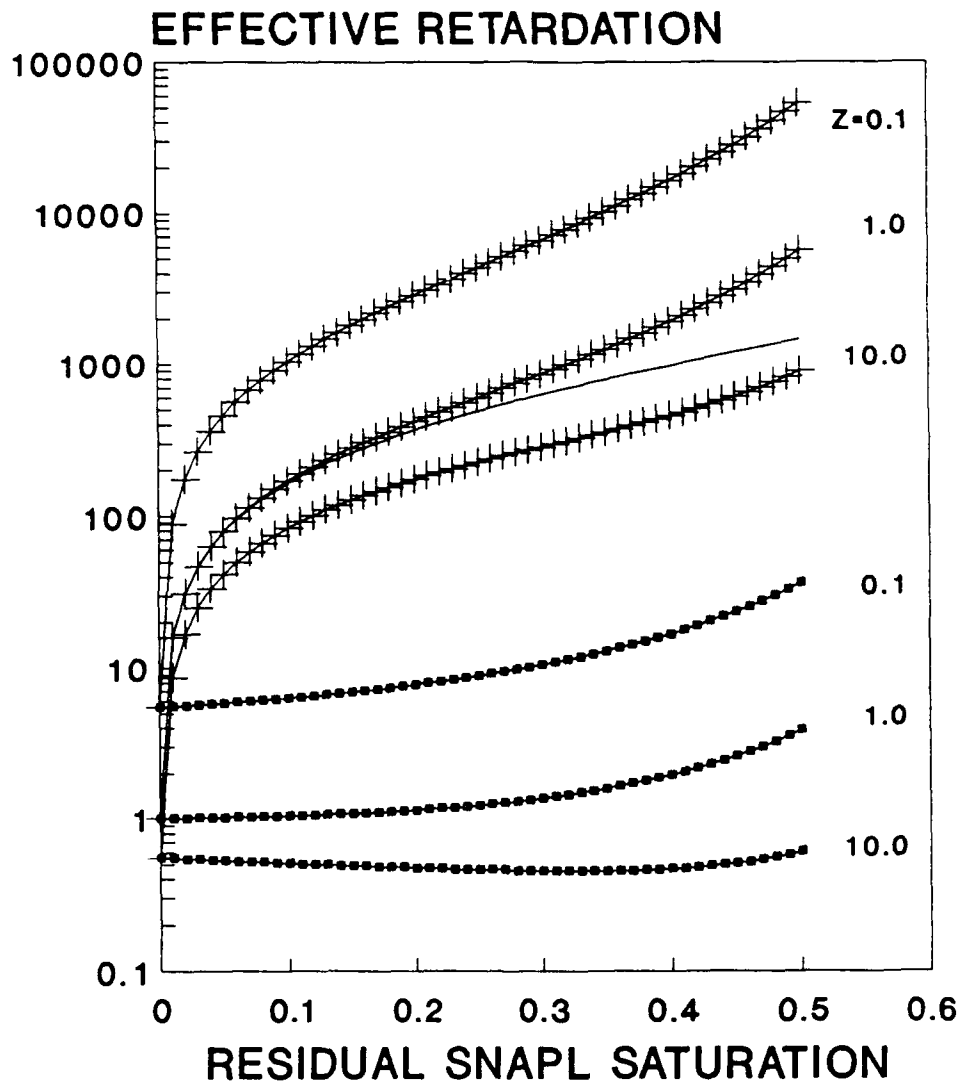


Fig. 11 The Effective Retardation in a Cylindrical Hypothetical SSS
($Z = k_h/k_{SA}$; + = 1-methylnaphthalene; — = R_{SSS}/R_o ; and
⊠ = Nonreactive tracer).

presented in this paper serve as a general model for effective retardation at both high and low residual SNAPL saturations. A family of curves for both 1-methylnaphthalene and a nonreactive contaminant are shown; the nonreactive contaminant is any dissolved constituent that is nonsorbing and conservative. Values of a dimensionless parameter Z , appear in Fig. 11 in association with each effective retardation curve. Z represents the ratio of permeability in the SSS at zero residual SNAPL saturation to the natural permeability of the surrounding aquifer.

The nonreactive contaminant curves were developed using Eq. 55. Both sorption parameters K_p and K_{ow} were set to zero, and Eq. 54 was used in conjunction with an experimental value for the residual water saturation S_{wm} , to define the relative permeability k_r .

From the nonreactive curves, effective retardation is not greatly affected by changes in residual SNAPL saturations in the SSS. The small increases in the pore velocity, caused by decreasing water filled porosity, are being canceled by a slightly greater decreases in the flow caused by decreasing SSS permeability. The parameter which is most important is the permeability ratio Z . When Z equaled 0.1, the permeability of the soil in the SSS was 10 times less than that of the surrounding aquifer; streamlines diverged as they approached the SSS and pore velocities decreased below ambient rates. The slower pore velocities produced effective retardations which ranged from 5.5 to 6.8 for residual SNAPL saturation less than 13.9 percent. For a SSS constructed with a soil having a permeability 10 times greater than the SA, streamlines converged near the sorption field, indicating an acceleration of flow toward the SSS (this is the case depicted in Fig. 10). The accelerated flow gave an effective retardation that was less than one, meaning nonreactive pollutants moved faster after SSS installation.

The effective retardation curves for 1-methylnaphthalene illustrate a potentially significant contaminant velocity reduction through the SSS. These curves were obtained in the same manner as described above, except that K_p equaled an experimental value of 11100 l/kg and K_{ow} equaled 7413 l/kg for 1-methylnaphthalene (Sangster 1989). The fraction of natural organic matter in the soil was assumed to be 0.02 percent for the SSS and

the surrounding aquifer. For a given value of Z , the reactive contaminant curves originated from the same starting points as the nonreactive curves; however, 1-methylnaphthalene sorption increased rapidly with residual SNAPL saturation. The increase in sorption caused reactive effective retardation curves to diverge rapidly from that of nonreactive constituents. In the range of observed decane/mineral oil residuals, sorption alone could potentially increase retardation of methylnaphthalene two orders of magnitude above a nonreactive constituent under the same flow conditions. Estimated effective retardations ranged as high as 1630 when Z equaled 0.1 and as low as 120 when Z equaled 10.0. For the latter case of Z equal to 10, the velocity of nonreactive constituents would increase as groundwater flow converges upon the SSS; nevertheless, the migration rate of 1-methylnaphthalene would decrease by a factor of 120 because of the SSS.

A first order approximation of the SSS capacity to retain contaminants may be estimated in terms of the system's effective retardation. This translates ambient contaminant velocities to what they will be once the sorption system is installed. An estimate of the time needed to exhaust a SSS can be obtained from

$$T_{0.5} = \frac{L_e R_{e|SSS}}{V_{co}} \quad (56)$$

$T_{0.5}$ is the time needed to achieve at the down gradient edge of the sorption field, a contaminant concentration which is one half of a constant resident concentration imposed at the entrance of the SSS. L_e is an equivalent travel length through the SSS. Sufficient radius would be needed to ensure a contaminant plume was intercepted.

In the past, engineers may have been tempted to avoid the complexity of the above analysis and estimate effective retardation by taking the ratio of the retardation factor inside the system R_{SSS} to that for ambient conditions R_o . This approach ignores the retardation caused by spatial changes in permeability; hence, ignoring the changes in the velocity field that result after the SSS is installed. The retardation factor for the ambient condition is found by substituting Eq. 43 into Eq. 42. R_{SSS} is

obtained after Eqs. 43, 44, 45, and 46 are combined with Eq. 42. The ratio of the retardation factors is plotted in Fig. 10. It now becomes evident, that the ratio of retardation factors is an inappropriate substitute for Eq. 55, unless the permeability and water filled porosity of the soil in the SSS is close to that of the surrounding media.

In the domain outside the boundary of the sorption system the effective retardation is obtained from the resultant specific discharge for flow in this region. The specific discharge components of \bar{q} in the SA are found from complex differentiation of Eq. 51b.

$$\frac{dN_{SA}}{dz} = -q_{x0} + q_{x0} \left(\frac{k_{SA} - k_{SSS}}{k_{SA} + k_{SSS}} \right) \frac{R_s^2}{(x^2 + y^2)^2} [x^2 + y^2 - 2xyi] \quad (57)$$

The real portion of Eq. 57 equals the specific discharge component along the x axis; thus,

$$q_x = q_{x0} \left[1 - \left(\frac{k_{SA} - k_{SSS}}{k_{SA} + k_{SSS}} \right) \frac{R_s^2}{(x^2 + y^2)} \right] \quad (58)$$

The imaginary part gives

$$q_y = -q_{x0} \left(\frac{k_{SA} - k_{SSS}}{k_{SA} + k_{SSS}} \right) \frac{R_s^2 xy}{(x^2 + y^2)^2} \quad (59)$$

The resultant specific discharge \bar{q} in the SA is then

$$\bar{q} = q_{x0} \left\{ \left[1 - \left(\frac{k_{SA} - k_{SSS}}{k_{SA} + k_{SSS}} \right) \frac{R_s^2}{(x^2 + y^2)} \right]^2 + \left[\left(\frac{k_{SA} - k_{SSS}}{k_{SA} + k_{SSS}} \right) \frac{R_s^2 xy}{(x^2 + y^2)^2} \right]^2 \right\}^{1/2} \quad (60)$$

Eq. 60 is valid for any combination of x and y where $(x^2 + y^2) \geq R_s$. In contrast to the region inside the sorption system, the specific discharge is not uniform throughout the domain of the SA. It is evident,

however, that at large distances from the sorption system, the specific discharge approaches that of the uniform flow specified in the far field.

An analytical expression for the effective retardation in the SA is found from combining Eqs. 47, 48, 49, and 60; the expression is simply the reciprocal of the resultant specific discharge in the SA normalized to the uniform flow rate from the ambient condition

$$Re|_{SA} = \left\{ \left[1 - \left(\frac{k_{SA} - k_r k_h}{k_{SA} + k_r k_h} \right) \frac{R_s^2}{(x^2 + y^2)} \right]^2 + \left[\left(\frac{k_{SA} - k_r k_h}{k_{SA} + k_r k_h} \right) \frac{R_s^2 xy}{(x^2 + y^2)^2} \right]^2 \right\}^{-1/2} \quad (61)$$

Note that the permeability of the sorption system has been substituted for the product of k_r and k_h .

A family of hypothetical effective retardation curves were developed for the point location created by the intersection of the upgradient boundary of the SSS and the x axis (see Fig. 12). Eq. 54 was substituted into Eq. 61 and solved for a broad range of residual SNAPL saturations. The general retardation effects of the permeability ratio Z , are the same as in the sorption zone. After deriving the effective retardation function for the SA, it becomes obvious that changes in apparent contaminant velocity are due entirely to the deviations from the ambient uniform flow field. These deviations are the result of permeability differences between the SSS and the SA alone. At locations further away from the SSS the hydraulic disturbances created by the system decrease and the effective retardation approaches 1.0. The curves in Fig. 12 apply to both reactive and nonreactive contaminants. The sorptive characteristics of the porous media are not altered in the SA; consequently, sorption should not affect estimates of effective retardation.

EFFECTIVE RETARDATION IN SURROUNDING AQUIFER

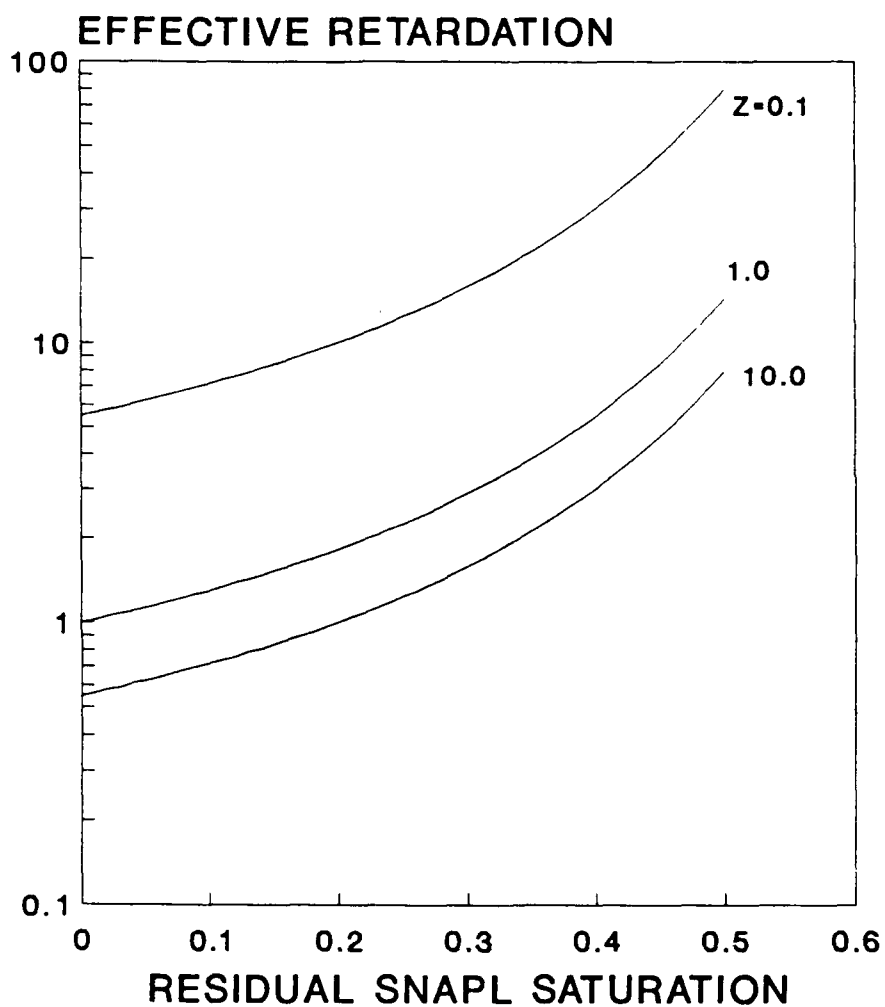


Fig. 12 Effective Retardation in the Aquifer Surrounding a Hypothetical Cylindrical SSS ($Z = k_H/k_{SA}$).

5.0 CONCLUSIONS AND RECOMMENDATIONS

Experimental and theoretical research has been conducted on subsurface contaminant sorption systems. The goal of the project was to develop a knowledge base that would be useful in designing sorption systems and predicting system performance in the field. In an effort to achieve this goal, the first research objective was to perform experiments with specific organics, that would deliver data needed to predict sorption system capacity and performance for those organics. The second objective was to provide a theoretical foundations leading to the development of analytical expressions which predict the performance of subsurface sorption systems with cylindrical geometry.

With regard to the first objective, experiments were successfully executed. Measured BTCs were clearly simulated using available nonequilibrium transport models. Basic parameters such as the retardation factor were accurately and independently estimated without the use of miscible displacement experiments.

With regard to the second objective, a theoretical analysis of a general cylindrical SSS produced a new dimensionless design parameter, the 'Effective Retardation'. The new parameter represents a concise expression of system performance in terms of residual SNAPL saturations, salient sorption parameters, and appurtenant permeabilities. An application, using experimental data on 1-methylnaphthalene, showed that a SSS could conceivably decrease plume velocities two to three orders-of-magnitude. Under typical hydraulic gradients, plumes could be retarded for 3 to 30 years.

With regard to future research, three major recommendations are forwarded

- 1) Independent methods of estimating nonequilibrium model parameters should be a research priority; but, only if subsurface sorption systems are likely to be used under induced groundwater flow conditions. Otherwise, natural gradients are not expected to produce the level of nonequilibrium that could result in premature contaminant breakthrough.

- 2) This research effort derived expressions for effective retardation in cylindrical SSSs. Theoretical work is needed to develop effective retardation expressions for other practical geometric configurations (i.e., rectangle, ellipse, crescent).
- 3) Laboratory and theoretical tools are now available to design and predict the performance of a cylindrical system. It is now appropriate to construct and test a system on an intermediate scale.

6.0 REFERENCES

- Banerjee, P., Piwoni, M. D., and Ebeid, K. (1985). "Sorption of organic contaminants to a low carbon subsurface core." Chemosphere. 14(8), 1057-1067.
- Bouchard, D. C., and Wood, A. L. (1988). "Pesticide sorption on geological material of varying organic carbon content." Toxic. Industr. Health. 4, 341-349.
- Burris, D., and Antworth, C. (1990). "Potential for subsurface in situ sorbent systems." Groundwater Management. 4, 527-538.
- Cameron, D. A., and Klute, A. (1977). "Convective-dispersive solute transport with a combined equilibrium and kinetic adsorption model." Water Resour. Res., 13, 197-199.
- Coats, K. H., and Smith, B. D. (1964). "Dead-end pore volume and dispersion in porous media." Soc. Pet. Eng. J., 4, 73-84.
- Enfield, C. G., Carsel, R. F., and Cohen, S. Z. (1982). "Approximating pollutant transport to ground water." Ground Water. 20(6), 711-722.
- Hamaker, J. W. (1975). "The interpretation of soil leaching experiments," in Environmental dynamics of pesticides. Haque, R., and Freed, V. H., Editors, Plenum Press, New York, 6, 115-133.
- Hatfield, K. (1989). "Contaminant flux reductions through in situ solubility modification." 1989 USAF-UES Summer Faculty Research Program: Final Report.

- Hatfield, K., Burris, D., Stauffer, T. B., and Ziegler, J. (1991). "Experimental and theoretical investigations of subsurface contaminant sorption systems." Submitted to the J. Envir. Engrg. Div., for review.
- Karickhoff, S. W., Brown, D. S., and Scott, T. A. (1979). "Sorption of hydrophobic pollutants on natural sediments." Water Resour. (13), 241-248.
- Lee, J., Crum, J. R., and Boyd, S. A. (1989). "Enhanced retention of organic contaminant by soils exchanged with organic cations." Environ. Sci. and Technol. 23(11), 1365-1372.
- MacIntyre, W. G., and Stauffer, T. B. (1988). "Liquid chromatography application for determination of sorption on aquifer material." Chemosphere. 17(11), 2161-2173.
- Parker, J. C., and van Genuchten, M. (1984). "Determining transport parameters from laboratory and field tracer experiments." Bulletin 84-3, Virginia Agricultural Experiment Station, Virginia Polytechnic Institute and State University.
- Rose, W. (1949). "Theoretical generalizations leading to the evaluation of relative permeability." Trans. AIME. 179.
- Sangster, J. (1989). "Octanol-water partition coefficients of simple organic compounds." J. Phys. Chem. Ref. Data. 18(3), 1111-1142.
- Schwarzenbach, R. P., and Westall, J. (1981). "Transport of nonpolar organic compounds from surface water to groundwater. Laboratory sorption studies." Environ. Sci. Technol. 15(11), 1360-1367.
- Selim, H. M., Davidson, J. M., and Mansell R. S. (1976). "Evaluation of a two-site adsorption-desorption model for describing solute transport in soil." Proceedings of the Computer Simulation Conference, Am. Inst. of Chem. Eng. Washington, D.C., pp. 444-448.
- Stauffer, T. B., MacIntyre, W. G. and Wickman, D. C. (1989). "Sorption of nonpolar organic chemicals on low-carbon-content aquifer materials." Envir. Toxi. and Chem. 8, 845-852.

- Strack, O. D. L., and Haitjema, H. M. (1981). "Modeling double aquifer flow using a comprehensive potential and distributed singularities. 2. Solution for inhomogeneous permeabilities." Water Resour. Res. 17(5), 1551-1560.
- Valocchi, A. J., (1985). "Validity of the local equilibrium assumption for modeling sorbing solute transport through homogeneous soils." Water Resour. Res. 21(6), 808-820.
- van Genuchten, M. Th., and Wierenga, P. J. (1976). "Mass transfer studies in sorbing porous media, 1, Analytical solutions." Soil Sci. Soc. Am. J. 40, 473-480.

Effects of Surfactants on Partitioning of Hazardous Organic Components of
JP-4 Onto Low Organic Carbon Soils

Dr. Kim F. Hayes, Assistant Professor
Department of Civil and Environmental Engineering
University of Michigan
Ann Arbor, Michigan 48109-2125

Final Report
Research Initiation Program
Universal Energy Systems
4401 Dayton-Xenia Road
Dayton, Ohio 45432

Sponsored by the U.S. Air Force Office of Scientific Research

May 1, 1991

Effects of Surfactants on Partitioning of Hazardous Organic Components of JP-4 Onto Low Organic Carbon Soils

by

Dr. Kim F. Hayes

ABSTRACT

The major objectives of this research initiation project were to demonstrate that the partitioning of polyaromatic hydrocarbons (PAHs) to soils can be enhanced by surfactant sorption, and that PAH affinity to soil depends on both the amount and structure of sorbed surfactant. The affinity of naphthalene for cetyl trimethyl ammonium bromide (CTAB)-coated silica was investigated and found to depend on the surface concentration of CTAB. A model of structure of the CTAB-surface coating as a function of coverage is reported that accounts for the naphthalene partitioning behavior. At the lower coverages, CTAB is sorbed primarily as hemimicelles, but as the amount of CTAB on the surface increases, surface micelles begin to form in increasing number. Since the surface micelles form a more hydrophobic structure, naphthalene molecules have a greater affinity for them compared to the hemimicelles. At high enough coverages, surfactant coatings are formed which have a greater affinity for hydrophobic contaminants than natural organic matter. The results of this study indicate that the structure and orientation of surfactant coatings can have a significant impact on the affinity of hydrophobic organic contaminants for mineral surfaces.

ACKNOWLEDGEMENTS

I thank the Air Force Systems Command, Air Force Office of Scientific Research, Universal Energy Systems and the Air Force Engineering and Services Center for their sponsorship and administration of this research.

Certain individuals should be mentioned by name for their contributions to making this research possible. In particular, I am indebted to Drs. Dan A. Stone , Thomas Stauffer and David Burris for stimulating discussions about the research needs of the Air Force which led to the development of the research initiation proposal which has supported this work. The careful experimental work and help with the preparation of this final report by Will Siegfried is also gratefully acknowledged.

TABLE OF CONTENTS

I. INTRODUCTION.....	6
II. STATEMENT OF THE PROBLEM	6
III. OBJECTIVES.....	6
IV. BACKGROUND	7
V. EXPERIMENTAL PROTOCOL.....	9
VI. RESULTS	11
VII. DISCUSSION	15
VIII. CONCLUSION.....	17
IX. FUTURE WORK	18
X. REFERENCES	18
XII. APPENDIX (Notation)	20

TABLE AND FIGURES

Table I. CTAB surface coverages and K_d values.....	12
Figure 1. Adsorption density and dynamic electrophoretic mobility for CTAB on 0.5-10 μm silica at 0.01M NaCl and pH 8.....	8
Figure 2. Structures of the surfactant adsorption isotherm. (After Stratton-Crawley and Shergold, 1981.)	9
Figure 3. Naphthalene isotherms on CTAB-coated silica at pH = 9.8.....	13
Figure 4. Naphthalene isotherms on CTAB-coated silica at pH = 8.1.....	13
Figure 5. Naphthalene sorption to silica plotted versus f_{OC}	14
Figure 6. Naphthalene sorption to silica plotted versus f_{SC} (assuming monolayer coverage and 26 $\text{\AA}^2/\text{molecule}$ (Stratton-Crawley and Shergold, 1981))......	14
Figure 7. Dynamic electrophoretic mobility as a function of the fraction of the silica surface covered by CTAB (assuming monolayer coverage and 26 $\text{\AA}^2/\text{molecule}$).	15

I. INTRODUCTION

Containment and cleanup of groundwater contaminated by jet fuel spills at US Air Force bases is an area of growing concern. One way to control pollutant migration or to help remove contaminants from spill sites is to introduce surfactants (surface active agents) into groundwater aquifers. These compounds can change the chemical properties of the interfacial regions between the fuel, water, and aquifer solids and change the mobility of both water-soluble and insoluble components of jet fuel. A better understanding of the chemical interaction of surfactants with aquifer material is required before surfactants can be most effectively used to contain contaminant plumes or clean up contaminated sites.

II. STATEMENT OF THE PROBLEM

The potential for medium-weight (2- and 3-ring compounds) polycyclic aromatic hydrocarbons (PAHs) to contaminate groundwater in soils with low organic carbon and low clay content is of primary concern to the Air Force and has been identified as such in a recent Broad Agency Announcement (BAA 90-002). Soil venting schemes currently being applied by the Air Force to JP-4 fuel spills result in a low volatility residue containing medium-weight PAHs which have moderate solubility. At JP-4 spill sites there is concern that these soluble organic contaminants will move rapidly out of biologically active zones and into the groundwater. One possible method to reduce the transport of these types of contaminants is to add surfactants to increase the organic carbon content of aquifer materials, thereby increasing sorption and reducing mobility. The Air Force is interested in evaluating the potential for surfactants to reduce PAH mobility at JP-4 spill sites.

III. OBJECTIVES

The major objectives of this mini-grant work were to demonstrate that the partitioning of PAHs to soils can be enhanced by surfactant sorption and that PAH affinity to soil depends on both the amount and structure of sorbed surfactant. Two types of measurements were made on a surfactant/mineral system: (1) surfactant and PAH sorption affinity measurements as a function of pH to find optimal conditions for PAH sorption, and (2) interfacial potential measurements to determine how the structure of sorbed surfactant affects subsequent PAH sorption.

IV. BACKGROUND

When a fuel product is accidentally introduced into the soil environment, its fate and transport depend on the physical and chemical properties of the soil and the chemical characteristics of the fuel. A particularly important property of a soil matrix in this context is its organic carbon content.

The capacity of a soil to sorb nonionic, hydrophobic, organic compounds is related to the organic carbon content of the soil (Chiou et al., 1979, 1983). Cationic surfactants can be added to clays to increase their sorption capacity, leading to enhanced sorption of low molecular weight hydrophobic compounds (Bouchard et al., 1988; Boyd et al., 1988). Some recent work has been conducted illustrating that adsorption of the nonionic surfactant polyethylene glycol can lead to enhanced PAH sorption onto aquifer solids (Podoll et al., 1988).

While these studies illustrate the potential for surfactants to affect hydrophobic contaminant sorption, none have attempted to relate the structure of the sorbed surfactant to contaminant affinity. In monitoring surfactant sorption, the most significant advances in understanding sorption behavior have been made by combining information from several types of measurements. In past studies, adsorption isotherms, electrokinetic data and contact angle measurements have yielded the greatest insights into structure of sorbed surfactants (Ginn, 1970; Fuerstenau, 1957; Fuerstenau et al., 1964; Somasundaran and Fuerstenau, 1966). Recently, fluorescence spectroscopic studies (Parcher et al., 1989) have been performed to clarify the role of an organic coating's structure in the sorption affinity of hydrophobic contaminants to organic coatings. Using a fluorescing hydrophobic probe molecule, Parcher et al. (1989) found that silica coated with a short chain (C_2) hydrocarbon and having a relatively lower organic carbon content (5.6%) was more effective at removing a hydrophobic solute than silica coated with longer chain (C_8 and C_{18}) hydrocarbons and having higher organic content (14% and 22%). Although, conventional wisdom suggests that the surface coated with a longer chain hydrocarbon and having a higher organic content would be more hydrophobic and should have a higher affinity for a sorbing hydrophobic compound, this was contrary to what was observed. Parcher et al. (1989) found that the structure of the organic moiety at the surface was responsible for the observed partitioning behavior, not the chain length of the organic coating or its percent by weight of the solid. Only by using methods that can identify the structure and orientation of surfactant molecules in the interfacial region can it be learned what types of surface organic coatings have the greatest affinity for various classes of hydrophobic contaminants.

A technique which has been used in the past and is used in this study to characterize the structure

of sorbed surfactants is the measurement of electrophoretic mobility as a function of surfactant sorption density. In particular, shifts in the magnitude and sign of the electrophoretic mobility as a function of surfactant coverages can be used to infer the structures of sorbed complexes (Somasundaran and Fuerstenau, 1966; Shergold, 1986; Yap et al., 1981). The approach is illustrated using electrophoretic mobility and CTAB (cetyl trimethyl ammonium bromide) sorption density on silica from data of Hayes (1991) (Figure 1).

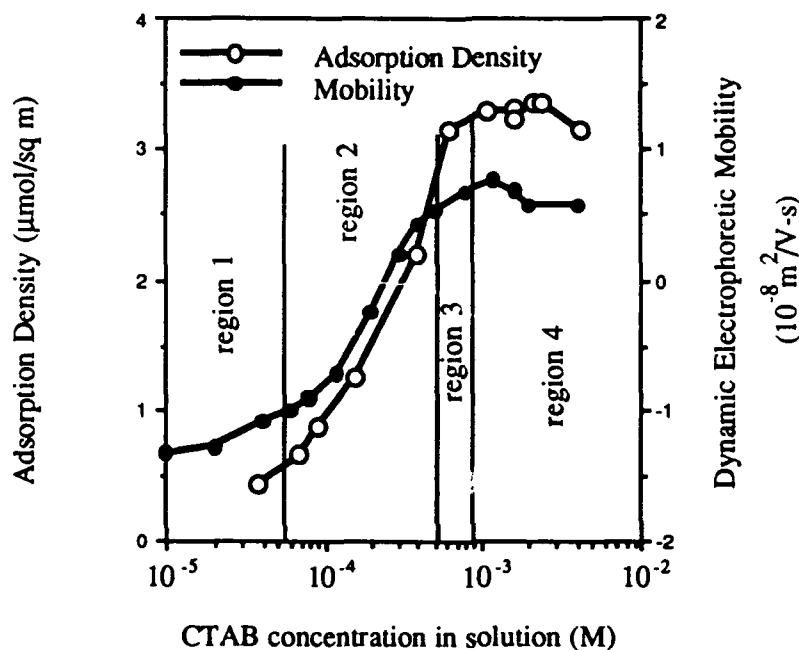


Figure 1. Adsorption density and dynamic electrophoretic mobility for CTAB on 0.5 - 10 μm silica at 0.01M NaCl and pH 8.

The CTAB sorption isotherm can be divided into the four regions shown in Figure 1. In region I, isolated CTAB molecules are thought to be sorbed by an electrostatic attraction between negatively charged surface sites and the positively charged surfactant. In this region, the silica surface is hydrophilic since it is still composed of mostly deprotonated surface hydroxyl groups as indicated by the negative value of the electrophoretic mobility. The marked increase in sorption in region II is thought to be mainly due to hemimicelle (cluster of surfactant molecules on the surface) formation, a result of lateral interactions between the hydrocarbon chains. The electrophoretic mobility is negative in the absence of CTAB, becomes less negative with the addition of CTAB, and finally changes sign. As the CTAB concentration increases further, the slope in the isotherm decreases (region III) reflecting a lower affinity of a positively charged surfactant for a positive surface. The increase in sorption in this region is thought to be mainly due to the favorable

hydrophobic interactions between the hydrocarbon portions of the surfactant molecules. This hydrophobic interaction may also lead to bilayer formation, which is thought to be negligible in region II, but substantial in region III. Finally, the sorption isotherm reaches a plateau (region IV) near the critical micelle concentration (CMC) of CTAB (9.0×10^{-4} M). The sorption density and the electrophoretic mobility remain constant at this point because the formation of solution phase micelles is energetically more favorable than additional sorption. Figure 2 schematically illustrates this interpretation of the structure and configuration of CTAB molecules on silica as a function of surface coverage.

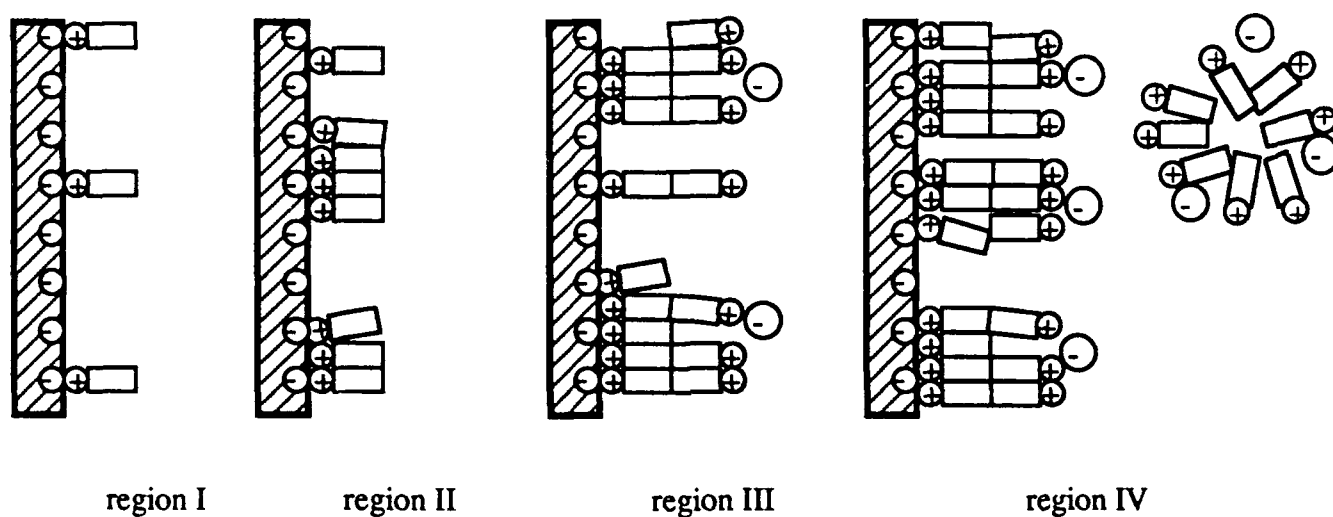


Figure 2. Structures of the surfactant adsorption isotherm. (After Stratton-Crawley and Shergold, 1981.)

In the present work naphthalene, an organic contaminant, is sorbed to CTAB-coated silica to investigate the effects of the orientation of surfactant molecules on their affinity for naphthalene. Previous solution phase work (Edwards et al., 1991) has demonstrated that the aqueous phase solubility of naphthalene is significantly enhanced at surfactant concentrations above the CMC. It is possible that the surface equivalent of a micelle in solution may also play a role in enhancing sorption of polyaromatic hydrophobic organic contaminants. As mentioned above, one of the goals of this study was to determine if a particular structure of sorbed surfactant would substantially enhance naphthalene partitioning.

V. EXPERIMENTAL PROTOCOL

Silica. The silica used in these experiments was #40 Sil-co-sil ground silica obtained from the U.S. Silica Company. This silica was from their Ottawa, IL mine, and was used without further treatment. Using BET adsorption, this silica was found to have a specific surface area of 5.6 m²/g.

Surfactant Adsorption Measurements. Batch sorption experiments were conducted using CTAB. The quantity sorbed was determined by the loss of solute from the liquid phase. The reactors used were 4 ml (nominal) glass vials with Teflon-faced rubber septa. Water, silica, surfactant (and its ¹⁴C radiolabelled tracer), background electrolyte (NaCl), and base (NaOH) were added in appropriate amounts to the vials. The ionic strength was kept constant at 0.01 M. Experiments were conducted at three different surfactant concentrations, 0.0583, 0.583, and 1.67 mM and at pH values of 8.1 and 9.8. Mixing of the vials was accomplished by end-over-end rotation of the reaction vials for a period of 16 - 24 hours at 8 rpm.

Solid/liquid separation was accomplished by centrifuging for two hours at 4,000 rpm at 20 °C on a Sorvall RC-5B refrigerated centrifuge with a GSA rotor (maximum relative centrifugal force = 2,604). For each sample, the pH was measured using a Ross combination pH electrode and an aliquot of supernatant was removed to measure ¹⁴C activity on an LKB Wallac 1219 Rackbeta liquid scintillation counter. In order to insure low surfactant solution concentration (well below the CMC) in the aqueous phase, most of the remaining supernatant was removed and replaced with an equivalent volume of surfactant-free "wash" water at 0.01M NaCl and the appropriate pH. The washing solution pH was adjusted to the pH measured prior to solid/liquid separation, and this was taken as the pH of the sample. The final surface concentration was determined by accounting for any additional desorption which occurred following this washing.

Naphthalene Sorption Measurements. Separate naphthalene sorption experiments were conducted without ¹⁴C labelled CTAB by replacing the washing solution described above with a solution of naphthalene having the desired pH, naphthalene concentration and an ionic strength of 0.01M (NaCl). The vials were filled to eliminate head space and then capped with Teflon lined rubber septa. The solids were resuspended by vortex mixing and then put on a rotator for 12 - 18 hours to allow equilibration. Then the vials were vortexed again (mainly to make sure there was no solid sticking to the caps) and centrifuged as in the CTAB sorption experiments.

Naphthalene was measured by injection of the supernatant into a Waters HPLC using a 70/30 methanol/water mobile phase and a Supelco 15 cm Econosphere 5µm C₁₈H₃₈ column. The

apparatus had a Lambda-Max Model 48 LC Spectrophotometer and an M-45 Solvent Delivery System operated at 1.0 ml/min.

Dynamic Electrophoretic Mobility Measurements. Dynamic electrophoretic mobility was measured using an electrokinetic sonic amplitude (ESA) measurement apparatus. In order to make ESA measurements, an alternating current is applied to a suspension of particles, causing these particles to undergo oscillatory motions due to the oscillating electrical forces on the charged particles. An acoustic wave develops due to the difference in density between the particles and the liquid. Pulsed ultrasonic methods with phase sensitive detection are used to measure the magnitude and phase angle of the electro-acoustic signal generated. The dynamic electrophoretic mobility is calculated from ESA measurements according to the theory described by O'Brien (1988). The instrument used for these measurements was a Matec ESA 8000 Measurement System.

VI. RESULTS

Figures 3 and 4 show isotherms for naphthalene sorbed to silica in both the presence and absence of a CTAB coating at two different pH values. The lines fitted to the data are used to calculate values of the naphthalene partition coefficient (K_d) for each coverage and pH combination measured. The partition coefficient, defined by

$$K_d = \frac{q_e}{C_e} \equiv \left(\frac{\text{moles/m}^2}{\text{moles/L}^3} \right) \equiv (\text{L/m}^2) \quad (1)$$

where q_e , C_e are the amount of naphthalene sorbed and equilibrium solution concentration, respectively, is the slope of each line in Figures 3 and 4. The values of the partition coefficients estimated for each coverage and pH value are presented in Table I. These K_d values can also be normalized for the fraction of organic carbon (f_{OC}), or for the fraction of the surface covered (f_{SC}) by the CTAB (assuming 26 \AA^2 per CTAB molecule (Stratton-Crawley and Shergold, 1981) and monolayer coverage) to give a sorption coefficient which is typically referred to as the K_{OC} value, e.g.,

$$K_{OC} = \frac{K_d}{f_{OC}} \quad (2)$$

where f_{OC} = (mass of surfactant sorbed per mass of silica) or

$$K_{OC} = \frac{K_d}{f_{SC}} \quad (3)$$

where f_{SC} = (m^2 of surfactant sorbed per m^2 of silica)

Table I. CTAB surface coverages and K_d values.

CTAB surface coverage	K_d	K_{OC}	pH
(mol/ m^2)	(L/ m^2)	(cm^3/g)	
1.49×10^{-7}	5.53×10^{-6}	1,311	8.1
1.44×10^{-6}	4.15×10^{-4}	1,985	8.1
3.26×10^{-6}	2.34×10^{-3}	2,976	8.1
1.81×10^{-7}	5.42×10^{-5}	163	9.8
1.66×10^{-6}	7.52×10^{-4}	1,262	9.8
3.44×10^{-6}	2.34×10^{-3}	3,151	9.8

Values of K_d normalized by f_{OC} and f_{SC} plotted versus f_{OC} and f_{SC} are presented as Figures 5 and 6, respectively. If the amount of organic carbon is the most important parameter in determining naphthalene partitioning, then values of K_{OC} would be expected to be independent of the surface coverage, e.g. K_{OC} would be constant. If, however, the structure and orientation of a surfactant is important in naphthalene partitioning, then the normalized values of K_d would be expected to change as a function of surface structure for naphthalene. Based on the results in Figures 5 and 6, which show that the value of K_{OC} increases with surfactant coverage, it is clear that the partitioning of naphthalene becomes more favorable at higher surfactant coverages. The values of K_{OC} in units of cm^3/g as a function of surfactant are also given in Table I.

To further clarify the role of structure and orientation of the surfactant on naphthalene partitioning as a function of surfactant surface coverage, the electrophoretic mobility of silica-coated particles was estimated by making ESA measurements. The results of the dynamic electrophoretic mobility are shown in Figure 7. As discussed above, the electrophoretic mobility measurements, in combination with the sorption isotherms data, can be used to infer the structure and orientation of sorbed CTAB (Figures 1 and 2). This information can then be used, in turn, to explain the structural basis for the enhanced affinity of naphthalene for CTAB-coated silica at high CTAB coverages.

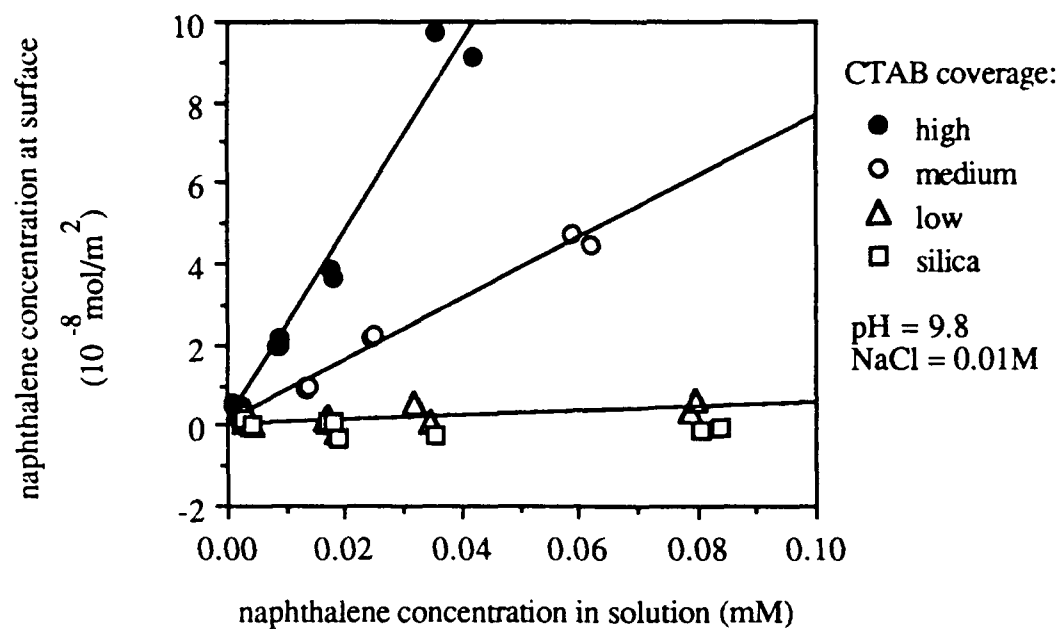


Figure 3. Naphthalene isotherms on CTAB-coated silica at pH = 9.8.

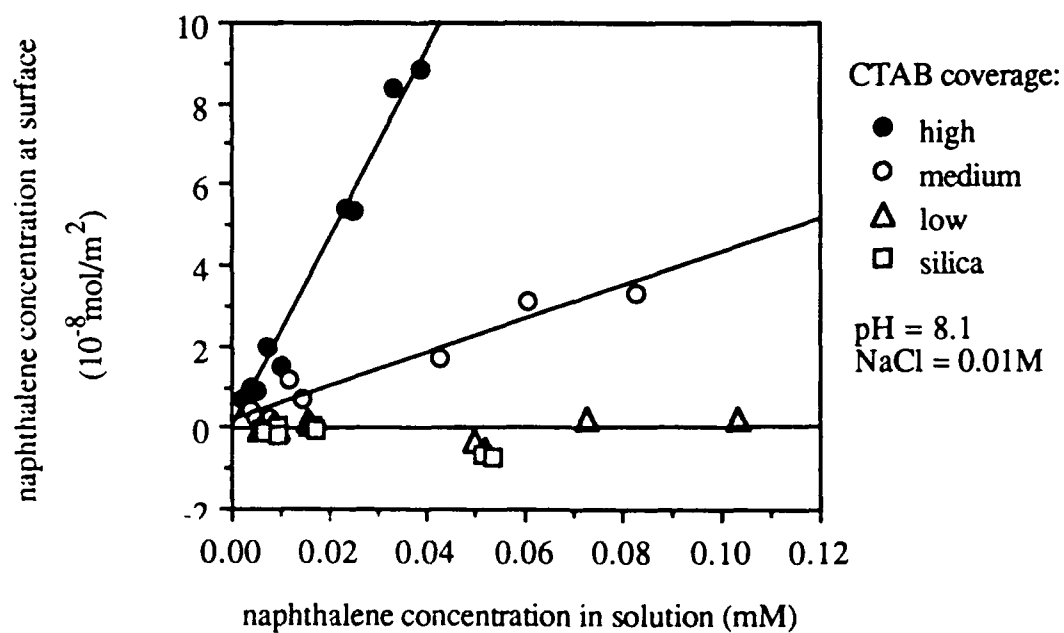


Figure 4. Naphthalene isotherms on CTAB-coated silica at pH = 8.1.

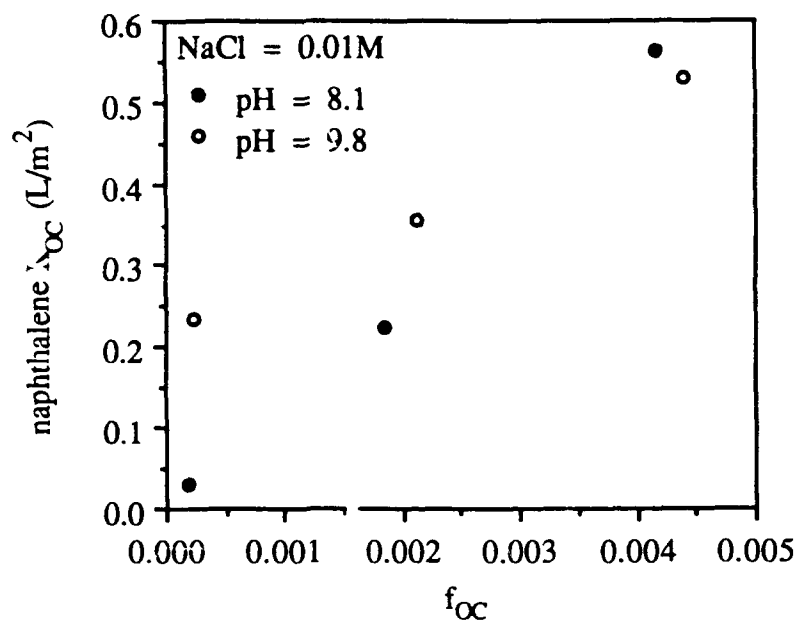


Figure 5. Naphthalene sorption to silica plotted versus f_{OC} .

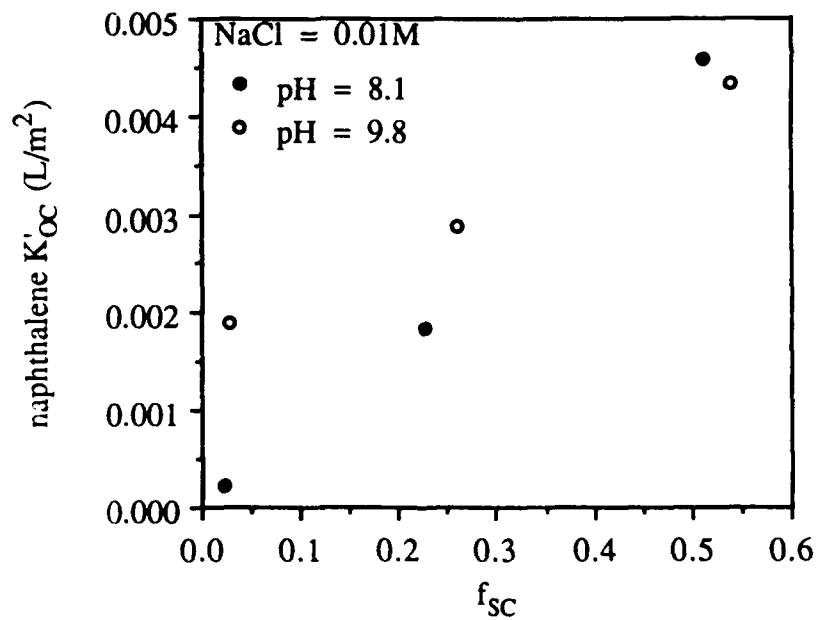


Figure 6. Naphthalene sorption to silica plotted versus f_{SC} (assuming monolayer coverage and 26 Å²/molecule (Stratton-Crawley and Shergold, 1981)).

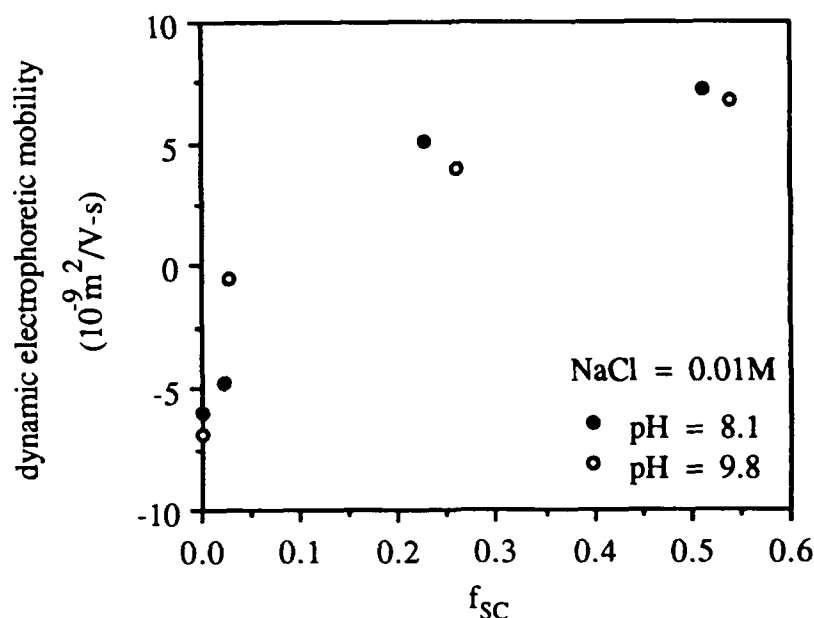


Figure 7. Dynamic electrophoretic mobility as a function of the fraction of the silica surface covered by CTAB (assuming monolayer coverage and $26 \text{ \AA}^2/\text{molecule}$).

VII. DISCUSSION

As shown in Figures 3 and 4, the naphthalene partitioning to uncoated silica particles is negligibly small. This is expected, since the surface of uncoated silica in water is essentially hydrophilic and should have little affinity for a hydrophobic solute like naphthalene, while the surfactant coating renders the surface more hydrophobic. As also shown, naphthalene sorption increases with increasing concentration of sorbed surfactant. This, too, is expected, because as the amount of CTAB on the surface increases, there is an increasing amount of an organic carbon phase on the surface to which naphthalene can partition. When the partition coefficient of naphthalene sorbed per mass of organic carbon, K_{OC} , is plotted versus surface coverage, it is found that K_{OC} also increases with coverage (Figures 5 and 6). In general, increasing values of K_{OC} with coverage are not expected, unless, as a function of surfactant coverage, the surface is becoming increasingly more hydrophobic per unit mass of organic carbon. Thus, the increase in K_{OC} observed with increasing surfactant coverage must be related to changes in the structure and orientation of the surfactant molecules on the silica surface as the surface coverage increases.

A plausible explanation of the increase in the value of K_{OC} with coverage can be made based on the description of the structure and orientation changes expected for sorbed CTAB as a function of coverage (Figures 1 and 2). Assuming that the low CTAB sorption coverages reported here are representative of hemimicelle formation of region II, and that the higher coverages are

representative of bilayer formation of region III, a plausible explanation of the increase of K_{OC} can be made. When naphthalene sorbs to a hemimicelle-like coating (region II), it is less effectively sheltered from the water phase and some of naphthalene molecule may still be exposed to the aqueous phase. At higher coverages, however, naphthalene is sorbed in the midst of the bilayer (region III). Since the bilayer region likely provides a more hydrophobic environment and may more completely shelter naphthalene from the water phase, it has a greater affinity for naphthalene, and hence it is expected that K_{OC} would increase as the relative concentration of the bilayer or surface "micelles" compared to the surface hemimicelles increases with coverage. That the organic-normalized partition coefficient, K_{OC} , increases monotonically with surfactant coverage indicates that the formation of surface micelles may be a continuous process, in contrast to the critical micelle concentration that occurs in the aqueous phase. However, more work is needed over a wider range of surfactant coverages and pH conditions to verify this.

Additional support for the above proposed structural model for surfactant sorption and naphthalene partitioning as a function of surfactant surface coverage, has been obtained from electrophoretic mobility measurements (Figure 7). As described above, surfactant sorption data, used in conjunction with electrophoretic mobility data, can be used to identify the structure and orientation of surfactants as a function of coverage (Figures 1 and 2). In particular, the transition from the hemimicelle to bilayer structure, in going from regions II and III, is thought to occur near the point of the reversal of the sign of the electrophoretic mobility as a function of surfactant coverage. For the conditions reported here, the electrophoretic mobility sign reversal occurs between the low ($f_{SC} = 0.02$) and medium ($f_{SC} = 0.20$) coverage ranges studied (Figure 7). This supports the conclusion that the sorption of naphthalene is occurring to CTAB-coated silica represented by regions II and III. Hence, this data provides further support for the contention that the surface is becoming increasingly more hydrophobic for naphthalene partitioning with coverage, with the transition from region II to region III occurring near the sign reversal of the dynamic electrophoretic mobility.

The effects of pH on the naphthalene partitioning also provide further support for the proposed model of a surface which becomes increasingly more hydrophobic with increasing surfactant coverage. As shown in Figures 5 and 6, K_{OC} for naphthalene increases monotonically at both the pH = 8.1 and pH = 9.8 samples. It is interesting to note that the slope of the K_{OC} versus f_{OC} is greater for the pH 8.1 compared to pH = 9.8 samples. A possible reason for this is that transition from hemimicelle to bilayer formation occurs at lower surfactant coverages at a pH value of 8.1 compared to pH 9.8 samples. Since, at the lower pH, uncoated silica has fewer negative sites to sorb surfactants in region I, it is possible that there are correspondingly fewer sites from which

hemimicelle and bilayers will form as coverage increases at lower pH. Hence, the transition from region II to III would be expected to occur at lower coverages and to give steeper slope values in a plot of K_{OC} versus f_{OC} at lower pH. Although a steeper slope is observed, the transition, based on the zero point of the electrophoretic mobility, does not appear to occur at lower coverages for the pH 8.1 compared to 9.8 systems (Figure 7). More work is needed on the effects of surfactant coverage, electrophoretic mobility, and naphthalene partitioning as a function of pH to verify this hypothesis.

Finally, a comparison of the values of K_{OC} for naphthalene sorption to surfactant coatings compared to natural soil organic coatings is instructive. An estimate of the value K_{OC} for soil organic carbon can be obtained from correlations that have been established between $\text{Log}K_{OC}$ and $\text{Log}K_{OW}$, the Log of the octanol/water partition coefficient (Karickhoff, 1984; Curtis et al., 1986). For example, Karickhoff has found the that relationship

$$\text{Log}K_{OC} = \text{Log}K_{OW} - 0.21 \quad (4)$$

describes the partitioning behavior of hydrophobic contaminants to natural organic matter quite well. Using the relationship of Eq. 4, and a value of $\text{Log}K_{OW}$ of 3.4 (Leo et al., 1971), the predicted value of K_{OC} is 1,550 (cm^3/g) which is near the values estimated from the lower- and mid-range coverage samples in this study (Table I). Hence, it can be seen that surfactant coatings on minerals can be generated which are even more effective than natural soil organic matter. This is not too surprising in view of the fact that surface micelles which are formed at higher coverage would be expected to be more hydrophobic than natural humic material coatings per mass of organic carbon. Others have reported similar enhanced partitioning of hydrophobic contaminants following surfactant modification of soils (Boyd et al., 1988).

VIII. CONCLUSION

The affinity of naphthalene for CTAB coated silica was found to depend on the surface concentration of CTAB. A model with a continuous transition between two surface orientations for the CTAB at the coverages reported here accounts for this behavior. CTAB at the lower coverages is sorbed primarily as hemimicelles, but as the amount of CTAB on the surface increases, it increasingly tends to form surface micelles. The surface micelles form a more hydrophobic structure, causing naphthalene molecules to have greater affinity for them compared to the hemimicelles. At high enough coverages, surfactant coatings can be formed which have a greater affinity for hydrophobic contaminants than natural organic matter.

IX. FUTURE WORK

This study indicates that the structure and orientation of surfactant coatings for mineral surfaces can have a significant impact on the affinity of hydrophobic organic contaminants for mineral surfaces. There are many interesting experiments which naturally follow from the results of this investigation. The results of the present study indicate that a continuous transition between a hemimicelle and micelle surface structure is occurring. However, the surface should eventually become predominantly surface micelles as coverage increases and the plot of K_{OC} versus f_{OC} should eventually level off when this happens. In addition, while our data seem to indicate that a continuous transition from a surface hemimicelle to micelle state is occurring as a function of coverage, additional work is needed to determine if there is, in fact, a critical surface micelle concentration, the equivalent of the critical micelle concentration in solution. These areas can be investigated by performing naphthalene sorption experiments at more surface coverages to determine, if over more narrow coverage ranges, a change in the slope of a plot of K_{OC} versus f_{OC} occurs. Since surface charge varies with pH, the importance of surface charge on the type of surface coating formed can be more thoroughly examined by expanding the investigation over a greater pH range. In addition, other surfactants and PAHs should be studied to determine if the conclusions drawn from this work are more generally applicable. Finally, it is also important to obtain more direct confirmation of the structure and orientation of sorbed surfactants as a function of surface coverages. We have recently found that by using self-supporting films and transmission Fourier Transform Infra-Red (FTIR) spectroscopy it is possible to investigate the structure of sorbed surfactants at mineral/water interfaces. Results from these preliminary FTIR studies are consistent with the description of the structure of surfactant surface coatings as a function of coverage as presented here. These FTIR results will be reported in the near future.

X. REFERENCES

- Bouchard, D.C., R.M. Powell, and D.A. Clark, "Organic Cation Effects on the Sorption of Metals and Neutral Organic Compounds on Aquifer Material," *J. Environmental Science and Health*, 1988, Vol. A23(6), pp. 585-601.
- Boyd, S.A., J.-F. Lee, and M.M. Mortland, "Attenuating Organic Contaminant Mobility by Soil Modification," *Nature*, 1988, Vol. 333, pp. 345-347.
- Chiou, C.T., L. Peters, and V.H. Freed, "A Physical Concept of Soil-Water Equilibria for Nonionic Organic Compounds," *Science*, 1979, Vol. 206, pp. 831-832.
- Chiou, C.T., P.E. Porter, and D.W. Schmedding, "Partition Equilibria of Nonionic Organic Compounds Between Soil Organic Matter and Water," *Environmental Science and Technology*, 1983, Vol. 18, pp. 295-297.

- Curtis, G.P., M. Reinhard, and P. Roberts, "Sorption of Hydrophobic Organic Compounds by Sediments," in *Geochemical Processes at Mineral Surfaces*, Eds. J. A. Davis and K.F. Hayes, ACS Symposium Series No. 323, Chapter, 10, 1986.
- Edwards, D.A., R.G. Luthy, and Z. Liu, "Solubilization of Polycyclic Aromatic Hydrocarbons in Micellar Nonionic Surfactant Solutions," *Environmental Science and Technology*, vol. 25, no. 1, Jan. 1991.
- Fuerstenau, D.W., "Correlation of Contact Angles, Adsorption Density, Zeta Potentials, and Flotation Rate", *Mining Engineering*, 1957, Vol. 9, pp. 1365-1367.
- Fuerstenau, D.W., T.W. Healy, and P. Somasundaran, "The Role of Hydrocarbon Chain Length of Alkyl Collectors in Flotation," *Transactions of the Society of Mining Engineers*, December, 1964.
- Ginn, M.E., "Adsorption of Cationic Surfactants on Mineral Substrates," in *Cationic Surfactants*, E. Jungermann, Ed., Marcel Dekker, Inc., New York, 1970.
- Hayes, K.F., unpublished data, 1991.
- Karickhoff, S.W., "Organic Pollutant Sorption in Aquatic Systems," *J. Hyd. Engng. Div. Am. Soc. Civ. Engrs.*, 1984, Vol. 110, 707-735.
- Leo, A., C. Hansch, and D. Elkins, "Partition Coefficients and Their Uses," *Chemical Reviews*, 1971, Vol. 71, 525-616.
- O'Brien, "Electro-acoustic effects in dilute suspension of spherical particles," *J. Fluid Mech.*, 1988, Vol. 190, 71-86.
- Parcher, J.F., C.J. Barbour, and R.W. Murray, "Fluorescence Study of Organic Cation Binding to Hydrocarbon-Bonded Silica Surfaces," *Analytical Chemistry*, 1989, Vol. 61, pp. 590-593.
- Podoll, R.T., and K.C. Irwin, "Sorption of Cationic Oligomers on Sediments," *Environmental Toxicology and Chemistry*, vol.7, pp. 405-415, 1988.
- Shergold H.L., "Cationic Surfactants in Mineral Processing," in *Industrial Applications of Surfactants*, D.R. Karsa, Ed., Royal Society of Chemistry symposium, 1986.
- Somasundaran, P., and D.W. Fuerstenau, "Mechanisms of Alkyl Sulfonate Adsorption at the Alumina-Water Interface," *The Journal of Physical Chemistry*, vol. 70, no.1, 1966.
- Stratton-Crawley, R. and H.L. Shergold, *Colloids and Surfaces*, 1981, Vol. 2, pp. 145-154.
- Yap S.N., R.K. Mishra, S. Raghavan, and D.W. Fuerstenau, "The Adsorption of Oleate from Aqueous Solution onto Hematite," in *Adsorption from Aqueous Solutions*, P.H. Tewari, Ed., ACS symposium, 1981.

XII. APPENDIX (Notation)

C_e	solution concentration in equilibrium with surface (mol/L)
CMC	critical micelle concentration
CTAB	cetyl trimethyl ammonium bromide; $C_{16}H_{33}N(CH_3)_3Br$
ESA	electrokinetic sonic amplitude
f_{OC}	fraction of organic carbon in system solids (g/g)
f_{SC}	fraction of surface coverage of system solids (m^2/m^2)
g	gram
K_{OC}	K_d/f_{OC} (L/m^2) or (cm^3/g)
K_d	partition coefficient, ratio of surface concentration to solution concentration (L/m^2)
L	liter
m	meter
ml	milliliter
mol	mole
PAH	polyaromatic hydrocarbon
q_e	surface concentration in equilibrium with solution (mol/m^2)

RESEARCH INITIATION PROGRAM

Sponsored by the
Air Force Office of Scientific Research

Conducted by the
Universal Energy Systems, Inc.

FINAL REPORT

Biodegradation of Hydrocarbon Components of Jet Fuel JP-4

Prepared by:	Deborah D. Ross, Ph.D.
Academic Rank:	Assistant Professor
Department:	Department of Biological Sciences
University:	Indiana Univ.-Purdue Univ.
Date:	31 December 1990

ABSTRACT

Removal of twenty hydrocarbons from a sandy loam soil was determined over a 15 day period. Within 5 days, all hydrocarbons with boiling points lower than that of undecane were removed from soil, whether untreated or treated with mercuric chloride to inhibit microbial activity. At 10 days, the long chain alkanes remained in the soil. Tridecane, tetradecane and pentadecane showed significantly greater removal in the untreated than in the treated soil, indicating that biodegradation played a role in removal of the higher boiling points hydrocarbons.

Microbial abundance and hydrocarbon-degrading ability were assessed in soil contaminated by jet fuel during a bioventing project. Soil samples from a contaminated site and an adjacent uncontaminated site were collected at the beginning and end of a 9-month project. Total direct counts, viable counts, and heterotrophic activity (radiolabeled glucose mineralization) were used to assess abundance and activity of the heterotrophic microbial population. Hydrocarbon-degrading capacity of the microbial population was assessed by estimating numbers of microorganisms capable of utilizing naphthalene, hexadecane and toluene (MPN method), and by determining the mineralization of the above three hydrocarbons in short (15 hour) and long-term (21-28 days) experiments. Results indicate that hydrocarbon contamination altered the activity of the bacterial population, producing lower numbers and heterotrophic activity of the population while increasing the hydrocarbon-degrading capacity of the population. By the end of the project, the activity of the bacterial population in the contaminated soil had shifted to a pattern characteristic of the uncontaminated site.

ACKNOWLEDGEMENTS

I wish to thank the Air Force Office of Scientific Research for sponsorship of this research and universal Energy Systems for administrative coordination of this program. I wish to thank Dr. Jim Spain and Dr. Charles Pettigrew for support, comments and suggestions.

INTRODUCTION

Previous studies on the environmental fate of JP-4 have focused on the aquatic environment. Biodegradation of the hydrocarbon components of JP-4 was studied by Spain and workers (1983). In these studies, JP-4 as well as a model fuel made up of known quantities of individual hydrocarbons present in JP-4 were added to water and sediment samples from several aquatic environments, and the disappearance of hydrocarbons was followed over several days. Use of aquatic samples which were killed by the addition of mercuric chloride allowed comparison of biotic and abiotic removal processes.

Results indicated that evaporation was the major removal process for the low molecular weight, volatile hydrocarbons. Addition of sediment to water samples affected the removal of JP-4 components by reducing the rate of volatilization. For most individual hydrocarbons, biodegradation was not as significant for removal as was evaporation. For those hydrocarbons which were susceptible to biodegradation, the extent of biodegradative removal was a function of the water sample. In general, water samples taken from environments which had seen previous exposure to hydrocarbons exhibited a greater extent of biodegradation than water samples taken from pristine environments.

To further elucidate the role of biodegradation in removal of JP-4, additional studies were conducted by Pritchard and coworkers (1988) using a quiescent bottle system designed to simulate a fuel spill on a quiet body of water. Results again indicated that volatilization was the primary removal process with lower molecular

weight hydrocarbons being lost at a more rapid rate than higher molecular weight hydrocarbons.

The primary goal of the recently completed study undertaken as part of the Summer Faculty Research Program was to evaluate the processes responsible for the removal of JP-8 from the environment. In addition to bottles receiving water and water/sediment slurries, a series of bottles containing soil were included in this study. The primary removal process in both water and water/sediment slurries was identified as evaporation. However, in the case of soil, biodegradation represented a significant removal process. For this reason, it is felt that parallel studies should be undertaken with JP-4 since comparable laboratory data on its biodegradative potential in soil are lacking.

The enhanced removal of hydrocarbons exhibited in soils containing active microbial populations raises the possibility of utilizing microorganisms in the remediation of contaminated soil. Bioremediation has received considerable attention as a cost effective means of removing petroleum (Bartha, 1986) as well as a wide variety of other organic chemicals (Wilson et al., 1986; Lee et al., 1988). These review articles record numerous examples of the removal of petroleum contaminants from soil. Most of these studies focus on the engineering design of the remediation technology and the degree of remediation achieved; in other words, bioremediation is treated as a "black box" into which contaminated soil is placed and out of which clean soil is obtained. Only a handful of studies address the microbial ecology of contaminated soils.

Thomas and coworkers (1989) studied microbial activity at a creosote waste site and demonstrated that while active microbial populations were present in pristine, slightly, and heavily contaminated soils, mineralization of polynuclear aromatic hydrocarbons was only significant in the contaminated soils. Thus, the microbial population has adapted to the presence of the contaminating hydrocarbons by increasing its biodegradative potential. A similar increase in numbers and activity of hydrocarbon degrading organisms was observed in a studies of microbial degradation of aromatic hydrocarbons at a hazardous waste site (Lee at al., 1984) and microbial activity in sediments receiving run-off from oil sand deposits (Wyndham and Costerton, 1981). In the latter case and in a study of phenol-degrading activity at an abandoned hazardous waste site (Dean-Ross, 1989), no correlation was observed between the extent of contamination and the activity of the microbial populations.

While the above studies provide information essential for an understanding of the microbial ecology of contaminated soils, they do not address the microbiology of bioremediation. None of these studies followed changes in microbial abundance and activity during an actual bioremediation project. The proposed study was designed to provide basic data on microbial abundance and activity to aid in the interpretation of engineering data from bioremediation projects.

OBJECTIVES

A. Overall objective: to evaluate the biodegradability of Air Force jet fuel JP-4 in contaminated soil.

B. Specific objectives:

1. To determine the rate and extent of biodegradation of JP-4 in soil.

Soil contamination of jet fuels from accidental spills or leaking underground storage tanks represents a recurrent input of hydrocarbons into the terrestrial environment. In a study conducted as part of the Summer Faculty Research Program, it was found that JP-8, a jet fuel used in Europe and proposed for use in the United States, was removed faster in soil containing an active population of microorganisms than in inactive soil, indicating that biodegradation is a significant factor in the environmental fate of jet fuels. Since no laboratory data is available for JP-4, the first objective will remedy this lack of data.

2. To determine the rate and extent of biodegradation of hydrocarbon components of JP-4 in uncontaminated soil, contaminated soil and soil contaminated with JP-4 but treated to encourage biodegradation.

The Environics Division of the Engineering and Services Center at Tyndall AFB will be initiating a field study entitled Enhanced Biodegradation through Soil Venting. This study will consist of four experimental plots, two contaminated by JP-4 and two uncontaminated by fuel. Moisture content, nutrient concentration and venting rate will be adjusted in the experimental plots to study the effect of these parameters on biodegradation. As part of this proposal, soil will be collected from

each of the experimental plots and biodegradation will be assessed in the laboratory by the measurement of release of radiolabeled carbon dioxide after the addition of representative radiolabeled hydrocarbons to the soil samples. It is proposed to use hexadecane, a representative straight chain alkane, and toluene and naphthalene, representative aromatic hydrocarbons.

3. To characterize the microorganisms responsible for mediating the biodegradation of JP-4.

The abundance and activity of microorganisms in the three experimental soils will be determined in order to assess the effect of treatment on the microbial population. In addition, microorganisms capable of utilizing each of the three model hydrocarbons will be isolated and characterized as to their biodegradative potential.

MATERIALS AND METHODS

A. Biodegradation of JP-4 in soil

Biodegradation of JP-4 was assessed using a procedure similar to that used for the assessment of JP-8 biodegradation. For this study, 25 g (dry weight equivalent) aliquots of soil were placed in 150 ml milk dilution bottles. Two sets of bottles were used, one set containing untreated (active) soil and the other receiving soil treated with 2% (wt) HgCl_2 . Each bottle received 250 μl of JP-4. Bottles were incubated in a horizontal position with caps removed. Triplicate bottles were removed at 0 time, and 5, 10, 20 and 30 days. Three bottles containing 25 g of soil (dry weight) were weighed and incubated under identical conditions to the JP-4 bottles. These bottles were weighed weekly to calculate weight loss. The corresponding amount of water was added to the JP-4 bottles to maintain a constant moisture level. Soil was extracted by addition of 25 mls of CS_2 , followed by shaking for 5 min on a wrist action mechanical shaker. Extracts were decanted from the soil, laced in vials and analyzed by high resolution capillary gas chromatography with a Perkin-Elmer Model 8500 gas chromatograph equipped with a flame ionization detector, using a Durabond 5 capillary column programmed to begin at 40° C with a 4 min isothermal period followed by 3°/min increase to 250° C.

B. Biodegradation of hydrocarbon components of JP-4 in experimental plots

To measure soil degradation of these hydrocarbons, a gas train similar to that employed by Ward (1985) was used. Soil was placed in Erlenmeyer flasks attached to a gas train supplying CO_2 free air at a controlled rate. Three series of flasks were

set up, one for each of three model hydrocarbons, toluene, hexadecane and naphthalene. JP-4 containing a spike of one of the above radiolabeled compounds was added to triplicate samples of soil from each experimental plot. As the air exited the flask, it passed through a Tenax AC trap (Heitkamp et al., 1987) to collect volatilized hydrocarbon followed by a CO₂ trap. Both traps were sampled at intervals over the experimental period and the amount of radioactivity in each measured by liquid scintillation counting.

C. Characterization of microbial populations in soils from bioventing project plots.

At initiation of the soil venting project (September, 1989) and at termination of the project (February, 1989), soil samples from the four experimental plots were collected, transported to IPFW and analyzed for microbial activity by the following methods:

1. Direct counts using epifluorescence microscopy according to the method of Balkwill and Ghiorse (1985).

2. Viable counts using PYTG agar, dilute PYTG agar and soil extract agar as described in Balkwill and Ghiorse (1985).

3. Heterotrophic activity as measured by utilization of glucose or a mixture of amino acids. In this procedure, a ¹⁴C-labeled substrate was added to aliquots of soil from the experimental plots in 60 ml serum vials. Five replicate vials were prepared per treatment; two vials received 0.5 ml 2N H₂SO₄ to serve as killed cell controls, the other three assessed the activity of the microbial population. Each vial received the labeled substrate. Flasks were capped with rubber stoppers fitted with center wells containing 0.1 ml of 10N NaOH absorbed onto a filter paper wick. Vials were

incubated for an appropriate time interval. At the end of the incubation period, reaction in the active vials was stopped by the addition of 0.5 ml 2N H₂SO₄. Vials were incubated with shaking for an additional hour to ensure complete trapping of CO₂. Wicks were removed, placed in a scintillation cocktail (Budgetsolve, Research Products International), and counted by liquid scintillation counting (Beckman Model LS 9800).

4. Hydrocarbon mineralization was assessed in a similar fashion using each of the three model hydrocarbon substrates.

5. Numbers of hydrocarbon degraders were determined using the Most Probable Number (MPN) procedure as modified by Somerville et al (1985) for each of the three model hydrocarbon substrates.

RESULTS

A. Biodegradation of JP-4 in soil

Results of quiescent bottle tests using JP-4 in Chippola soil are shown in Table 1. Disappearance of twenty hydrocarbon components of JP-4 were followed over a two week period. For hydrocarbons having boiling points below that of undecane, essentially all of the hydrocarbon disappeared from flasks by day 5. Hydrocarbons with boiling points higher than undecane were removed, but although not to the same extent as the lower boiling point hydrocarbons. At day 5, no significant differences were noted when comparing disappearance in flasks containing soil treated with HgCl_2 as compared to untreated soil. However, at day 10, hydrocarbon concentrations for the longer chain alkanes were lower in untreated than treated soils, indicating that biodegradation did play a significant role in removal of these hydrocarbon components.

B. Biodegradation of model hydrocarbons in soil from treatment plots

Toluene was removed from soil flasks by evaporation as indicated by entrapment of radiolabeled toluene on the resin filters by the second sampling day. No radiolabel was recovered in the KOH traps.

In soil samples taken from treatment plots at the beginning of the bioremediation project, hexadecane was biodegraded in the soil from the uncontaminated site as indicated in Figure 1. No significant biodegradation was observed in soil taken from the contaminated site. Naphthalene was biodegraded in

both soils to approximately the same extent.

In soil samples taken from treatment plots at the end of the bioremediation project, biodegradation was observed only in the case of hexadecane in uncontaminated soil (Fig. 2).

C. Characterization of microbial populations in soils from treatment plots

At the beginning of bioremediation, viable counts, direct counts, numbers of hydrocarbon degraders and hydrocarbon degrading activity was higher in soil from the contaminated site than the uncontaminated site. Heterotrophic activity, however, was higher in the uncontaminated than the contaminated soil.

At the end of bioremediation, viable counts, direct counts and heterotrophic activity were higher in the uncontaminated than the contaminated soil. The number of hexadecane degraders was slightly higher in the uncontaminated than the contaminated soil. No naphthalene or toluene degraders were detected in uncontaminated soil. Hydrocarbon degrading activity was below the detection limit in both contaminated and uncontaminated soils.

RECOMMENDATIONS

1. Biodegradation is a factor in the removal of high boiling point hydrocarbon components of jet fuel. Bioremediation projects therefore have the highest chance of success in focusing on these hydrocarbons.

2. Biodegradation of hexadecane was lower in hydrocarbon contaminated soil than in uncontaminated soil, in spite of higher numbers of degraders. This suggests that environmental factors may be limiting biodegradation in field situations.

3. Hydrocarbon contamination enriches for a microbial population containing significant numbers of hydrocarbon degrading bacteria. Manipulation of this population to maximize the rate of biodegradation may be a cost effective means of enhancing removal of fuel from contaminated soil.

REFERENCES

- Balkwill, D. L. and W. C. Ghiorse. 1985. Characterization of subsurface bacteria associated with two shallow aquifers in Oklahoma. *Appl. Environ. Microbiol.* 50:580-588.
- Bartha, R. 1986. Biotechnology of petroleum pollutant biodegradation. *Microb. Ecol.* 12:155-172.
- Carlson, R. E. 1981. The biological degradation of spilled jet fuels: a literature review. USAF/ESC Report ESL-TR-81-50.
- Cook, A. M., H. Grossenbacher, and R. Hutter. 1983. Isolation and cultivation of microbes with biodegradative potential. *Experientia* 39:1191-1198.
- Haigler, B. E., S. F. Nishino, and J. C. Spain. 1988. Degradation of 1,2-dichlorobenzene by a Pseudomonas sp. *Appl. Environ. Microbiol.* 54:294-301.
- Heitkamp, M. A., J. P. Freeman and C. E. Cerniglia. 1987. Naphthalene biodegradation in environmental microcosms: estimates of degradation rates and characterization of metabolites. *Appl. Environ. Microbiol.* 53:129-136.
- Lee, M. D., J. M. Thomas, R. C. Borden, P. B. Bedient, C. H. Ward, and J. T.

Wilson. 1988. Biorestitution of aquifers contaminated with organic compounds. CRC Crit. Rev. Environ. Control 18:29-89.

Lee, M. D., J. T. Wilson, and C. H. Ward. 1984. Microbial degradation of selected aromatics in a hazardous waste site. Devel. Ind. Microbiol. 25:557-565.

Pritchard, P. H., T. P. Maziarz, L. H. Mueller, and A. W. Bourquin. 1988. Environmental fate and effects of shale-derived jet fuel. USAF/ESC Report ESL-TR-87-09.

Spain, J. C. and C. C. Somerville. 1985. Fate and toxicity of high density missile fuels RJ-5 and JP-9 in aquatic test systems. Chemosphere 14:239-248.

Spain, J. C., C. C. Somerville, L. C. Butler, T. J. Lee, and A. W. Bourquin. 1983. Degradation of jet fuel hydrocarbons by aquatic microbial communities. USAF/ESC Report ESL-TR-83-26.

Somerville, C. C., C. A. Monti, and J. C. Spain. 1985. Modification of the ^{14}C most-probable-number method for use with nonpolar and volatile substrates. Appl. Environ. Microbiol. 49:711-713.

Thomas, J. M., M. D. Lee, M. J. Scott and C. H. Ward. 1989. Microbial ecology of the subsurface at an abandoned creosote waste site. J. Ind. Microbiol. 4:109-120.

Ward, T. E. 1985. Characterizing the aerobic and anaerobic microbial activities in surface and subsurface soils. *Environ. Toxicol. Chem.* 4:727-737.

Wilson, J. T., L. E. Leach, M. Henson, and J. N. Jones. 1986. In situ bioremediation as a ground water remediation technique. *Ground Water Monitor. Review.* Fall, 1986:56-64.

TABLE 1. REMOVAL OF HYDROCARBONS FROM UNTREATED SOIL

Hydrocarbon	Concentration in extract (ug/ml) at:			
	O time	5 day	10 day	15 day
2,3-dimethylpentane	.061	0	0	0
heptane	.277	0	0	0
methylcyclohexane	.094	0	0	0
toluene	.087	0	0	0
3-methylheptane	.136	0	0	0
octane	.197	0	0	0
ethylbenzene	.060	0	0	0
m-xylene	.067	0	0	0
o-xylene	.034	0	0	0
nonane	.118	0	0	0
iso-butylbenzene	.028	0	0	0
1,3,5-trimethylbenzene	.035	0	0	0
1,2,4-trimethylbenzene	.068	0	0	0
decane	.102	0	0	0
undecane	.108	.018	.012	.011
dodecane	.111	.056	.016	.014
tridecane	.099	.067	.022	.018
tetradecane	.057	.067	.020	.019
pentadecane	.027	.032	.017	.016
hexadecane	.015	.010	.008	0

TABLE 2. REMOVAL OF HYDROCARBONS FROM TREATED SOIL

Hydrocarbon	Concentration in extract (ug/ml) at:			
	0 time	5 day	10 day	15 day
2,3-dimethylpentane	.050	0	0	0
heptane	.235	0	0	0
methylcyclohexane	.081	0	0	0
toluene	.091	0	0	0
3-methylheptane	.121	0	0	0
octane	.177	0	0	0
ethylbenzene	.057	0	0	0
m-xylene	.063	0	0	0
o-xylene	.033	0	0	0
nonane	.110	0	0	0
iso-butylbenzene	.027	0	0	0
1,3,5-trimethylbenzene	.035	0	0	0
1,2,4-trimethylbenzene	.057	0	0	0
decane	.095	0	0	0
undecane	.102	.018	.010	.006
dodecane	.104	.056	.015	.017
tridecane	.094	.070	.043	.029
tetradecane	.056	.065	.049	.029
pentadecane	.031	.038	.032	.018
hexadecane	.015	.010	.008	.005

TABLE 3. ABUNDANCE AND ACTIVITY OF THE MICROBIAL
COMMUNITY AT THE BEGINNING OF BIOREMEDIATION

	Soil Treatment	
	Contaminated	Uncontaminated
Viable counts (per g soil)	2.48×10^6	1.97×10^6
Direct counts (per g soil)	3.02×10^8	1.34×10^8
Hydrocarbon degraders (per g soil)		
Hexadecane	5.20×10^6	8.32×10^5
Naphthalene	8.32×10^5	8.32×10^5
Toluene	8.32×10^6	8.32×10^4
Heterotrophic activity (turnover time in hours)	96.0	51.6
Hydrocarbon-degrading activity (turnover time in hours)		
Hexadecane	ND	ND
Naphthalene	459.9	977.2
Toluene	2074	ND

ND = none detected

**TABLE 4. ABUNDANCE AND ACTIVITY OF THE MICROBIAL
COMMUNITY AT THE END OF BIOREMEDIATION**

	Soil Treatment	
	Contaminated	Uncontaminated
Viable counts (per g soil)	3.19×10^6	5.74×10^6
Direct counts (per g soil)	1.35×10^8	8.88×10^8
Hydrocarbon degraders (per g soil)		
Hexadecane	2.29×10^5	3.47×10^5
Naphthalene	5.2×10^3	ND
Toluene	5.2×10^4	ND
Heterotrophic activity (turnover time in hours)	217.0	9.40
Hydrocarbon-degrading activity (turnover time in hours)		
Hexadecane	ND	ND
Naphthalene	ND	ND
Toluene	ND	ND

ND = none detected

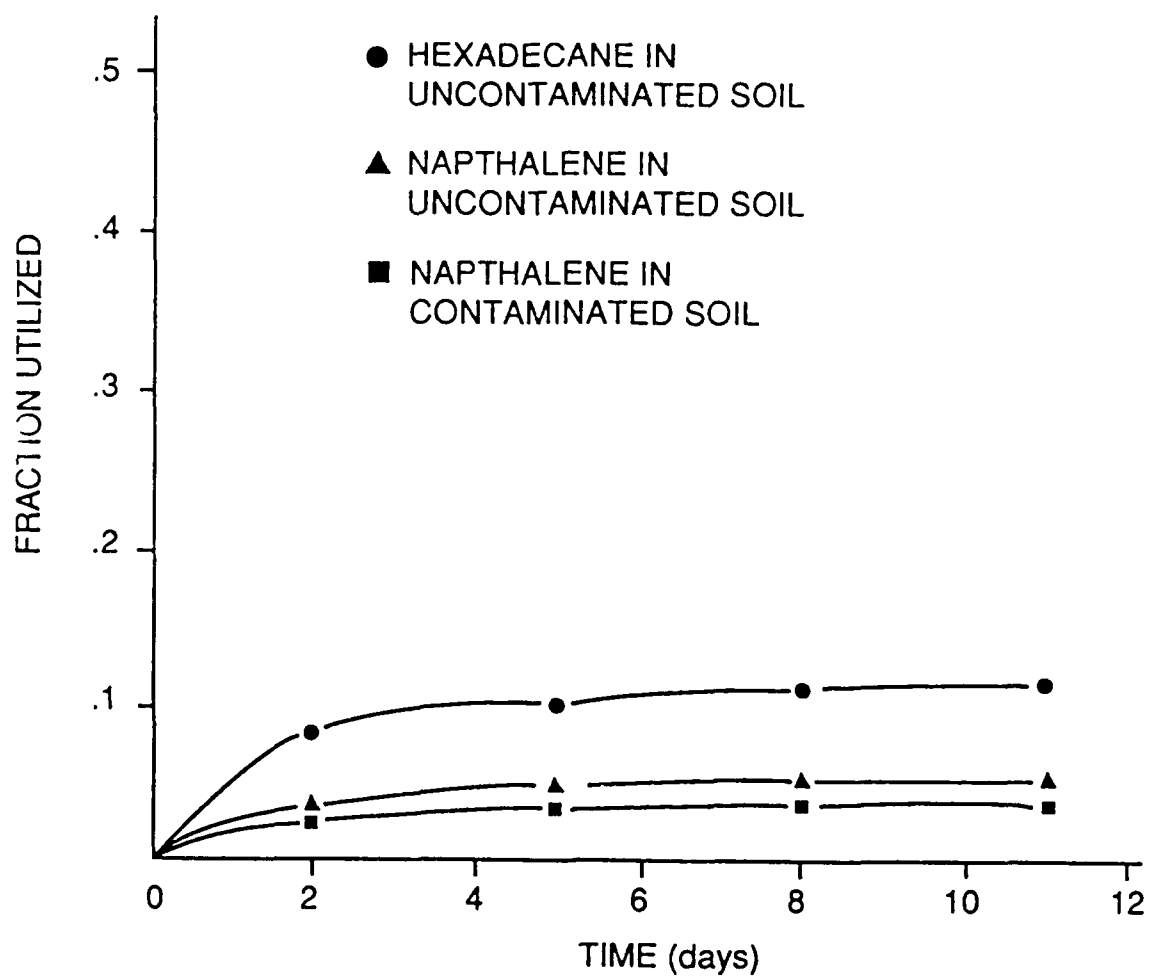


Figure 1. Hydrocarbon biodegradation at the beginning of bioremediation.

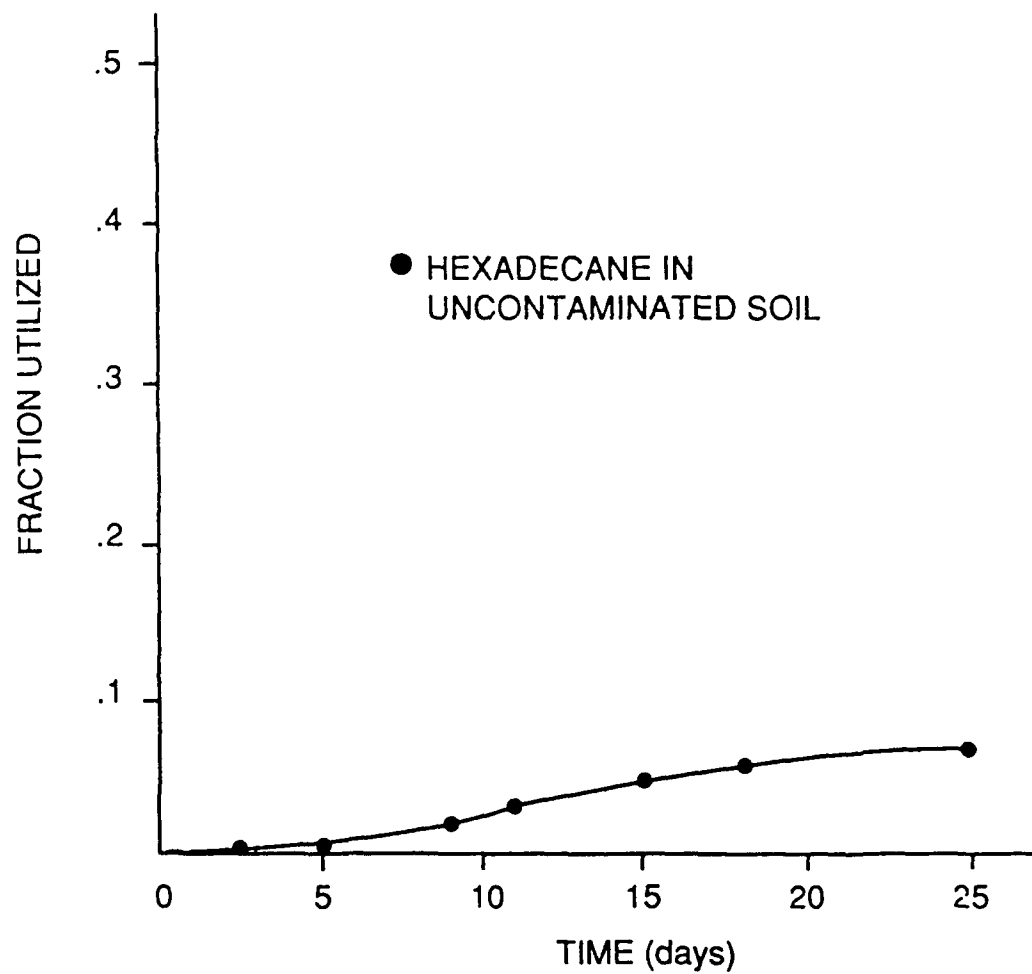


Figure 2. Hydrocarbon biodegradation at the end of bioremediation.

Report # 28
760-7MG-079
Prof. William Schulz
See Report # 71
210-10MG-018

1987-88 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the
Universal Energy Systems, Inc.

FINAL REPORT

PRETREATMENT OF WASTEWATERS GENERATED BY
FIREFIGHTER TRAINING FACILITIES

Principal Investigator:	Dennis D. Truax, Ph.D., P.E.
Academic Rank:	Assistant Professor
Department and University:	Department of Civil Engineering Mississippi State University
Address:	P. O. Drawer CE Mississippi State, MS 39762
USAF Effort Coordinator:	Mr. Bruce Nielsen Air Force Engineering and Services Center HQ AFESC/RDVW Tyndall AFB, FL 32403-6001
Effort Period:	15 JAN 88 through 15 JAN 89
Contract No.:	UES-S-760-7MB-105

ABSTRACT

Commercial airports and Air Force bases throughout the United States train personnel to extinguish fires associated with aircraft flight operations. A waste stream consisting of jet fuel, surfactant-based firefighting agents, and particulates is created by these necessary exercises. The Air Force is developing a treatment scheme for these wastes.

This project evaluated two pretreatment approaches for reducing foaming potential and enhancing biodegradability of the wastewater. First, chemically coagulation was examined as a way of surfactant removal because they are colloidal in nature. Aluminum sulfate and ferric chloride were studied as the primary coagulants. Chemical oxidation was the second approach investigated as a way to alter the foaming and bioinhibitory character of the surfactants. Hydrogen peroxide and potassium permanganate were used for this phase of the project. A synthetic waste and grab samples from three training facilities were evaluated for treatability by both approaches.

The results show that the surfactants in the wastewater can not be removed through the coagulation/flocculation process. However, an interesting foam inhibition effect was observed with the addition of the coagulant and adjustment of pH. In particular, if the pH was held to about 5.0 units, foam generation did not occur, but only if a small amount of alum had been previously added.

Regarding chemical oxidation, hydrogen peroxide proved of no value. However, the addition of potassium permanganate was found to permanently eliminate foam generation potential of the wastewaters evaluated. This was achieved at moderate doses of permanganate if the pH was lowered to a level of 6.5 units. However, the biodegradability and effective organic content of the samples was not significantly altered.

ACKNOWLEDGEMENTS

I would like to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of my research. In addition, I appreciate the help and cooperation of Universal Energy Systems during the project.

Thanks are also extended to the personnel at the Environics Laboratory of the Air Force Engineering Services Center (AFESC/RDVW). Special acknowledgement is needed for the technical help, support and patience provided by Mr. Bruce Nielsen. Through his efforts my experiences at Tyndall AFB have been pleasant ones. In addition, Lt Edward Marchand is to be thanked for his help in review and coordination during the final project period.

Finally, I need to acknowledge the efforts of two graduate assistants, Mr. Ethan Merrill and Mr. David Collins, without whose help this project would never have been completed.

PRETREATMENT OF WASTEWATERS GENERATED BY
FIREFIGHTER TRAINING FACILITIES

by

Dennis D. Truax, Ph.D., P.E.

INTRODUCTION

Training personnel in the techniques of dealing with aircraft fires is an important activity at airports. In addition to the numerous training facilities at commercially-operated airports, the Air Force has over one hundred firefighter training sites located at military airfields.

Recently the Fire Technology Branch (RDCF) of the Air Force Engineering and Services Center (AFESC) designed an environmentally secure firefighter training facility (FTF). In general, the FTF consists of a circular burn pit, a vehicle maneuvering area, fuel storage and supply system, smoke reduction apparatus, and a wastewater handling area. The burn pit has plastic and clay liners. These prevent the migration of liquid to the surrounding soil and underlying groundwater. After a training activity, the pit will contain a mixture of water, unburned aviation fuel (i.e., JP-4), soot particles, and a fire extinguishing agent consisting principally of aqueous film-forming foam (AFFF). [1-4] This combination of liquids and solids constitutes a wastewater which must be treated before release into the environment.

At present, the specifications for the wastewater handling system call for an oil/water separator. This unit was designed to reduce the

oil and grease content of the wastewater to or below 25 mg/l and allow recovery of unburned fuel.[1] At the same time, particulates will be removed to degree as yet undetermined. Therefore, the waste stream from the separator will consist primarily of water laden with AFFF and water-miscible hydrocarbons from the unburned fuel.

The aqueous effluent from the separator will then enter an impermeable holding basin designed in such a way to enhance evaporation. In the event that "...unfavorable meteorological conditions (high precipitation and low evaporation) cause the basin to exceed its design capacity...", the wastewater will have to be removed and undergo further treatment.[1] In that many regions experience rates of precipitation that are greater than the rate of evaporation, the Environics Division (RDVW) of AFESC is charged with evaluation of appropriate wastewater treatment schemes.[2]

Though the exact chemical composition of AFFF is proprietary and varies between manufacturers. It is known that AFFF consists of a mixture of fluorocarbon and non-fluorocarbon surfactants, solubilizers, and water. The surfactant molecules impart to the wastewater properties such as foaming, emulsification, and particle suspension.[5] Further, these chemicals inhibit biological activity when present in sufficient concentration.[3,6]

JP-4 is a classification of the jet fuel now utilized by the Air Force. It is the fuel used for Air Force FTF activities. JP-4 is composed of several potentially hazardous organics including the priority pollutants toluene, ethylbenzene, and naphthalene. However, the water-miscible components of JP-4 are primarily low molecular weight

cycloalkanes and aromatics. A large portion of these compounds are lost during FTF activities due to their low vapor pressures. Therefore, the fire training facility wastewater can be expected to predominantly contain higher molecular weight alkanes that normally have very low solubility values.[7]

Various treatment alternatives exist which might be able to handle these materials. These choices can be divided into two broad classifications, segregated and joint treatment. In segregated treatment, a separate and complete wastewater processing scheme would be developed to provide an effluent which could be discharge into the environment. Zachritz[2] and Chan[3] have examined various systems finding mixed results. Further, given the highly variable nature of waste stream flows and characteristics, consistent operation of segregated systems may be extremely difficult.

Joint treatment is the discharge of wastewater into a publicly owned treatment works (POTW) or the sewerage system of the authority operating the FTF; i.e., airbase sewer or treatment plant. Based on data from a recent survey[8], over forty percent of Air Force installations are using POTWs to treat their wastewaters. It is anticipated that virtually all of these facilities incorporate biological treatment processes which could be adversely impacted by the addition of untreated FTF wastewater. Of those airbases still operating treatment facilities, about eighty percent use biological processes. Of these, many are having trouble meeting treatment requirements because of age or design. The addition of untreated FTF wastewaters could aggravate this problem.

If joint treatment at POTWs is practiced, pretreatment of this wastewater will be required under current federal regulations. This would be due, in part, to the potential of these discharges to cause two problems. The first is excessive foaming in the treatment plant's reactors and effluent. The other is a reduction in effluent quality resulting from a bioinhibitory nature of the wastewaters components and/or overloading of biological reactors unacclimated to these components. Both problems have been observed during laboratory testing of wastewater samples from FTFs as well as investigations into its components. [3,4,6,9]

This study is an initial examination of selected pretreatment strategies for wastewaters generated by Air Force FTFs. It seeks to determine if chemical coagulation and chemical oxidation are viable pretreatment techniques for reducing foaming potential and increasing biodegradability of these wastewaters. Some work has been performed in this area. [3,9,10] However, much of the data collected for FTF wastes is dated. The chemicals and operational procedures were different from those used today. Some of the information reported seems contradictory. Finally, these projects did not thoroughly examine alternatives which might have application to pretreatment.

OBJECTIVES OF THE RESEARCH EFFORT

The primary goal of pretreating this wastewater is the reduction of the foaming potential. A secondary goal is improving the potential for bacterial conversion of the organics contained in it. The literature shows there are a number of alternatives for removing oils and surfactants.[11-13] These processes include chemical oxidation and chemical coagulation/flocculation. Because these processes are appropriate for treating the components of FTF wastewaters, it is logical to expect that they would have application in pretreatment of the waste stream. Previous research efforts would appear to support this.[3,9,10] This project will examine each of these in some detail.

BACKGROUND

AFFF is an effective fire suppressant due, in part, to the surface active agents contained in it. Surfactants are materials which dissolve, or tend to dissolve, in water and non-aqueous materials. A surfactant molecule contains a strongly hydrophobic group and a strongly hydrophilic one. Such molecules tend to congregate at interfaces between the aqueous medium and other phases of the system (e.g., air, oily liquids, particulates). They lower the surface tension of the water at these interfaces thus imparting properties such as foaming, emulsification, and particle suspension.[5] To reduce this effect, one must remove or chemical modify the surfactant in a way which eliminates its hydrophilic and/or hydrophobic property.[7, 14-16]

Chemical Coagulation

Surfactants are colloidal in nature. It is customary to distinguish two classes of colloids whose general behavior is entirely different. These classes are called lyophobic and lyophilic colloids. Lyophilic colloids have a strong affinity for the dispersing solvent while in lyophobic colloids this attraction is weak or absent. As discussed above, surfactants in water have both lyophilic and lyophobic components.

In most systems, colloids are held in suspension, or stabilized, because of electrostatic forces which they have, or develop, relative to the surrounding water.[17] In particular, colloids of similar nature

will develop similar charges. Bodies having similar charges repeal each other. Due to this repulsive forces, the colloids remain in suspension. Further, the polar nature of the water molecule results in an attraction of the colloid to the bulk liquid phase which retards separation.

The parameters governing flocculation of dispersed droplets of an emulsion and dispersed solid particles are the same.[15, 17, 18] The term stability, when applied to emulsions, refers to the resistance of the dispersed droplets to coalescence. The floating or settling of the droplets because of a difference in density between them and the continuous phase is not considered instability.

An important physical property of colloidal dispersions is their tendency to aggregate. Collisions between dispersed colloids occur frequently as a result of Brownian motion. Attractive forces, termed van der Walls' forces, exist between all colloidal particles regardless of dissimilar chemical natures.[19] These attractive forces are responsible for the colloidal aggregation. However, their magnitude depends upon the kinds of atoms which make up the colloidal particles and the density of the particles. Also, the attractive force decreases with distance separating the particles.

When considering the aggregation of colloids it is useful to distinguish between colloidal transport and colloidal destabilization. Colloidal transport is a physical process related to the probability that the colloid's kinetic energy is sufficient to overcome the repulsive electrostatic energy barrier. This is accomplished by phenomena such as Brownian diffusion, fluid motion, and sedimentation. It is controlled by such physical parameters as temperature, velocity

gradient, and colloidal size. The destabilization of colloidal suspensions is achieved by reducing the energy of interaction between two colloids and by formation of chemical bridges which aggregate colloids into three-dimensional floc networks.[19, 20]

The metal ions used for coagulation aid in the latter mechanism. Through considerations of ionic strength, the addition of highly charged metals reduces the energy of interaction. Then, as the metals hydrolyzed in water to form polynuclear complexes, they aggregate with the destabilized particles due to van der Waals forces. This process can be aided by agitation of the bulk liquid. This causes the particles to come in close vicinity and increases the chances that they will collide and coalesce.[17]

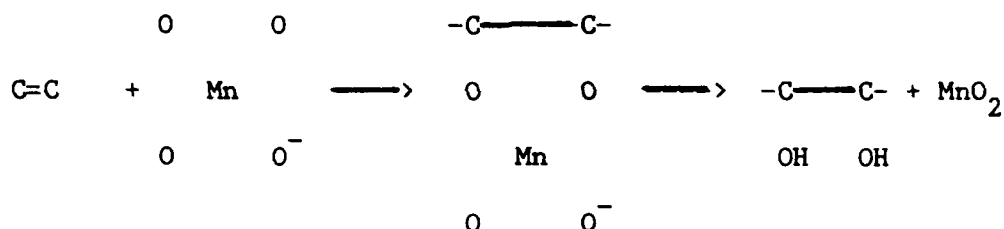
The most widely used coagulants in water and wastewater treatment are aluminum and iron salts. Aluminum sulfate is employed more frequently because it is usually cheaper. Iron has the advantages of forming a denser, rapid settling floc over a wider pH ranges.[17] The best pH range when using alum is 4.5 to 8.0 due to the low solubility of aluminum hydroxide within this range. For ferric chloride a broader pH range of 4 to 12 is optimal.

Chemical Oxidation

Ladbury and Cullis[22] suggest two principle mechanisms exist by which organic compounds are oxidized by acid permanganate. The first is direct oxidation by the permanganate ion. The second involves formation of an intermediate Mn^{4+} ion and subsequent oxidation by hydroxyl radicals produced from this ion and water. For direct oxidation to

proceed, the organic should have a hydrogen atom amenable to substitution and/or oxidation. No such site seems mandated if the Mn^{4+} intermediate provides oxidation because this processes exhibits a higher activation energy than direct oxidation. Though the kinetics of permanganate oxidation of organic compounds can be first-order initially, this relationship can not be maintained. In many instances second-order kinetics are exhibited throughout acid permanganate oxidation.

Probably the most common oxidative process associated with permanganate is syn hydroxylation of alkenes.[23, 24] This process produces glycols through the formation of cyclic intermediates followed by cleavage of the oxygen-metal bond as illustrated below:



Limiting the reaction to the hydroxylation reaction illustrated above is very difficult under normal conditions. Therefore, it is presumed that further oxidation of the glycol occurs resulting in the destruction of the organic molecule.

When dealing with FTF wastewaters, the most significant surfactant is of the linear alkyl sulfonate type. Typically, the aggregate surfactant material contains both a hydrocarbon and fluorohydrocarbon surfactant specie bonded to a glycol carrier.[21] Alkyl sulfonates having a branched alkyl residue are formed by direct sulfonation of naphthenic and branched paraffin hydrocarbons. It is suggested that the

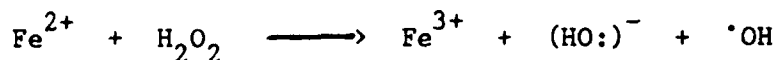
sulfonation of these tertiary olefins produces a product with a radical substitution at an activated methyl group while maintaining the integrity of the double bond as illustrated below:



The presence of the double bond as present in the product of this reaction provides a reactive site for direct permanganate oxidation.

Research has shown that oxidation of aromatic hydrocarbons in an acidic environment requires reduction of the permanganate ion to manganese dioxide and oxidation of the aromatic ring occurs concurrently with the side chain destruction.[22-24] Further, the extent of ring oxidation decreases as the size of the alkyl group in the monoalkylbenzenes increases. This facts imply that direct permanganate oxidation would be slow to react with fuel hydrocarbons.

Merz and Waters[25] examined the reaction of the free hydroxyl radicals produced by hydrogen peroxide acting on ferrous iron. They determined that the complete reaction between ferrous ion and hydrogen peroxide occurs in two steps as illustrated below.



and



They further found this radical, when in an acidified solution, capable of oxidizing water-soluble, aliphatic compounds through dehydrogenation. They indicate the dehydrogenation mechanism was followed by

chain and/or non-chain reactions which produce alkenes and alcohols in some cases and refractory products in others.

If the dehydrogenation-oxidation reactions occur, they can be sustained with a small amount of ferrous salt. This results in the destruction of large amounts of organic material with only a little catalyst required. If oxidation of the organic can not occur without catalyst being added, then the dehydrogenated radical stabilizes and the reaction stops.[25]

Finally, it is important to emphasize that these reactions must be performed under acidic conditions to prevent the ionization of the molecular hydrogen peroxide. Sims recommends an initial pH of 4.0 units and a 10 mg/L dose of ferrous ammonium sulfate (FAS) catalyst.[26]

Foam Measurement Techniques

Emulsions developed by combining surfactants and water have a high potential for foaming. When this solution enters an aeration basin, foam is produced. If this tank is a biological reactor at a POTW, operational problems develop; e.g., reduced aeration efficiency, floating solids, and slippery foam covering gantries and reactor systems.

A significant component of providing pretreatment of FTF waste streams is accurate quantification of foaming potential. Without a reproducible method of defining the foaminess of the waste, it is impossible to define treatment requirements or efficiencies. The literature presents a variety of foam measurement techniques. However, many of these are not suitable for this application due to method specificity or

FTF waste characteristics (e.g., suspended solids, turbidity, interfering complexes).

Bikerman's[14] work was aimed at establishing foaminess as a definite physical property of a liquid, which must therefore be independent of the apparatus used and the amount of material employed in measuring it. The approach he found most successful employed a dynamic foam meter. For this procedure, an inert gas is delivered at a measured rate through a porous membrane to an overlying layer of solution. The foam develops until a constant, maximum volume is reached.

Using this procedure, the average lifetime before bursting of a bubble in the foam describes the foaminess of the liquid. Designated as Sigma, this characteristic is determined using the following equation:

$$\text{Sigma} = \frac{V_o \cdot t}{V}$$

where: V_o is the maximum, steady-state volume of foam produced, t is the time it takes to reach V_o , and V is the total volume of gas used to generate the foam volume.

This approach is the basis of most recent work on foaming measurement. These efforts have examined procedural variables including geometry of the foam reactor, volume of solution sparged, type of sparger, and rate of gas delivery. This literature indicates that no single set-up will provide results for one water that are translatable to data for another. It also demonstrates that reproducible results can be obtained if the same surfactant(s) is measured, if gas flow rates are moderate to low, and if sidewall effects of the reactor are kept to a minimum.

GENERAL RESEARCH APPROACH

During this evaluation of FTF waste stream pretreatment, the primary constraint was reduction in the foaming potential of the wastewater. Enhancement of bacterial activity was viewed as a secondary criteria. This set of priorities was established for various reasons. First, it was felt there was a good probability the FTF waste would be treated at a POTW which incorporated aeration as a process. Secondly, extensive work has been performed to define limits on waste input to biological systems. It appears that, with proper control over the bacterial population and their environment, conventional POTWs can remove the organics from fire fighter training facility wastewaters. [27,28] On the other hand, development of foam from FTF wastes would disrupt normal operation of the POTW.

Samples

This project examined pretreatment of four wastewaters. The first, a synthetic waste, was used during initial experimentation due to the unavailability of actual waste samples. A water: fuel: AFFF concentrate ratio of 1: 0.225: 0.005 was selected for the synthetic waste. This ratio was felt representative of the relative quantities of each constituent used during an FTF exercise. [28] It was noted that during a training event the component ratio would change as constituent concentrations are reduced through combustion. However, it was felt

that using these volumes would produce a wastewater more difficult to treat and would therefore be conservative from a contaminate standpoint.

Sample development consisted of combining 15 liters of distilled water, 3.375 liters of JP-4 aircraft fuel, and 0.075 liters of Ansul 3% AFFF concentrate in a 24-liter glass carboy. Using a magnetic stir bar and base, the mixture was stirred continuously for 15 hours to promote hydrocarbon transfer to the water phase. After mixing, the liquid sat undisturbed for nine hours to allow for segregation of free fuel. The synthetic waste was siphoned from below the floating fuel-AFFF layer which developed, placed in 2.5-liter glass bottles, and stored in the dark at 4°C. Samples were warmed to room temperature immediately prior to testing and only in volumes sufficient for the work at hand.

In addition to the synthetic waste, grab samples were collected from three operational FTFs. One came from Columbus Air Force Base (CAFB), Columbus, Mississippi while two were obtained from different FTFs located at Tyndall Air Force Base (TAFB), Panama City, Florida. The facilities at TAFB consisted of an older, more traditional system (TAFB's old FTF) and a new, prototype training facility (TAFB's new FTF). Though new fire training facilities will use many of the design features of the TAFB prototype, this new FTF was expected to generate somewhat uncharacteristic wastes because of its use as a test bed for training and operational procedures and experimental chemicals.

Samples were obtained from lined holding ponds that follow oil/water separators for new TAFB FTF and CAFB's facility. The sample from TAFB's older facility was taken from the last stage of its oil/water separator in that no holding pond exists. In all cases,

samples were collected about four weeks after a training exercise and stored in glass carboys in the dark at 4°C. Only volumes of sample sufficient for a single experiment were warmed to room temperature immediately before testing.

Wastewater Characterization

Several different parameters were used in the characterization of the wastewaters tested and evaluation of process efficiencies.

Foam Measurement - Determining the degree of foaminess of the waste was the first obstacle overcome in the design of this experiment. The technique allowed comparative evaluation of treatment processes and initial waste characterization. For characterization, dilutions were analyzed to determine the initial strength of the four wastewaters. Each dilution was developed by combining distilled water with the sample. Dilution of the sample increased until reaching a concentration that resulted in no foam production.

Bikerman's[14] technique provided the basis upon which the procedure was developed, though it was modified for use in measuring AFFF. Bikerman examined ABS (alkylbenzene sulfonate) which is very different from AFFF. Through experimentation, it was determined that AFFF produces more foam than ABS under the same test conditions. The larger volume increased the sidewall effects for Bikerman's procedure to a point where large discontinuities in foam consistency, bubble size and foam collapse were observed. This yielded highly variable, almost unreproducible results for any given sample.

After further experimentation with glass and plastic tubes and chambers of differing diameters and geometries, a four-liter Nalgene graduated cylinder was selected. A wheatstone sparger was used to disperse the air as it was added to the bottom of the sample. A schematic of this system is illustrated in Figure 1.

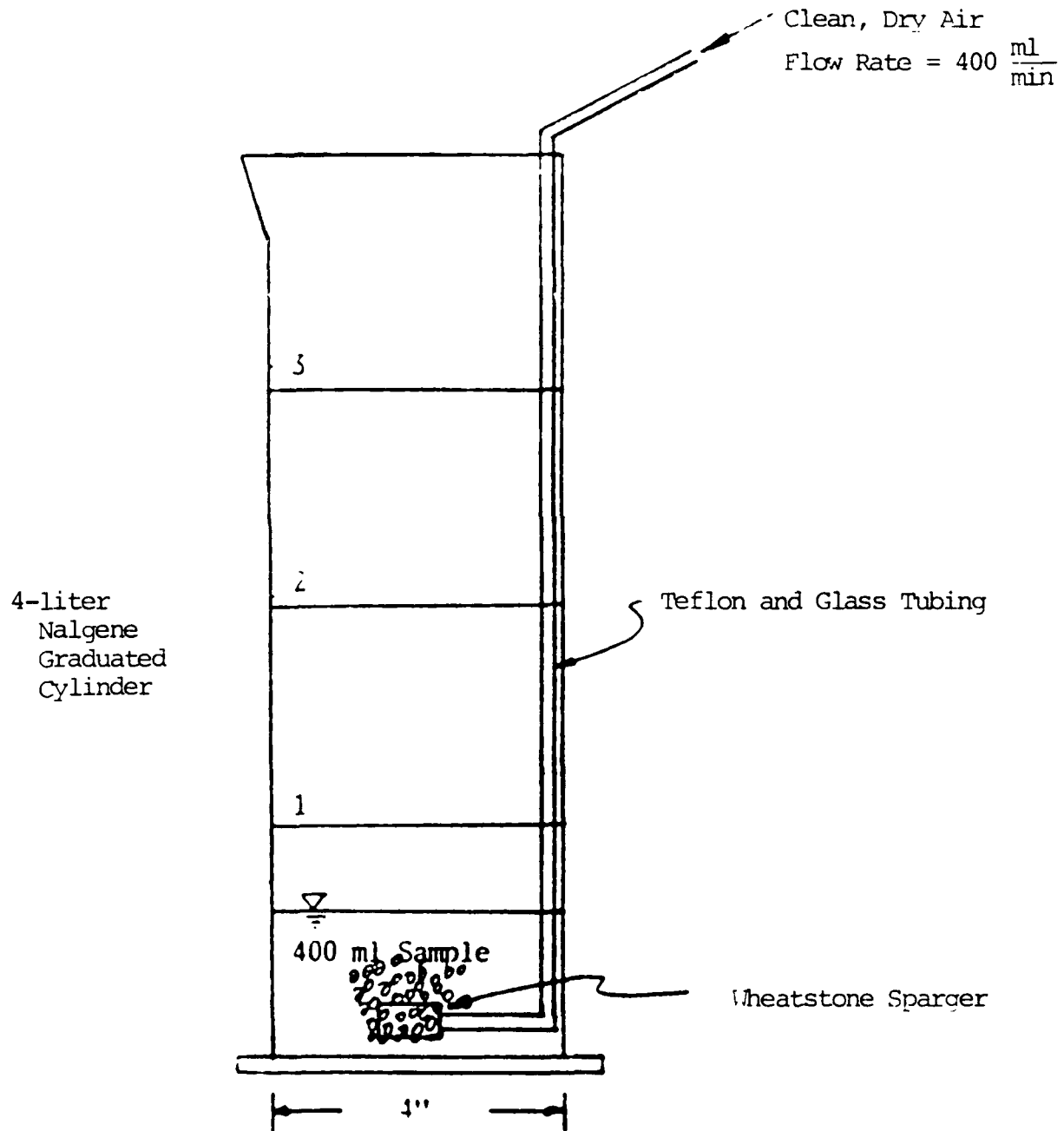
Four hundred milliliters of the sample were used. The sample was carefully metered down the wall of the graduated cylinder to prevent any foam production. After the sample had reached quiescent conditions, it was aerated at a constant rate of 200 ml/min for 4.0 minutes. For AFFF, the 200 ml/min air flow rate and 4-minute time interval showed excellent replicability and accuracy. At the end of the aeration period, the volume of foam generated was measured in milliliters and recorded.

Another deviation from Bikerman's technique regards the use of Sigma as a means of expressing the foaminess. By establishing a constant air flow rate over a constant time, the only variable in the Sigma relationship is that of the foam volume generated. Thus, Sigma did not need to be calculated. Rather, the volume of foam generated was recorded. It is felt that reporting the volume of foam generated makes it easier for the reader to perceive foaminess.

Bioinhibition Studies - Even if the aesthetic problems associated with this particular waste were neglected, the waste still could not be treated in a municipal treatment system without resolution of its bioinhibitory nature. To quantify the toxic threshold of the waste, thereby demonstrating the dilution needed for discharge to a POTW, a bioinhibition test was run on raw samples from each source.

Figure 1

Schematic of Foam Testing Apparatus



To determine this maximum concentration, the Biochemical Oxygen Demand (BOD) test was utilized.[2] Different volumes of each wastes were added to BOD bottles. A sample of an unacclimated bacterial culture was placed in the bottles. This bacterial seed was developed by collecting sludge from a local wastewater treatment plant and feeding the organisms sodium acetate as a substrate. Then, the bottles were filled with dilution water, initial dissolved oxygen measured, and capped. The one-day BOD was evaluated and results evaluated to determine when a significant reduction in bacterial activity occurred.

Biodegradability - The BOD test was used to indicate the biodegradability of the treated waste samples. The general equation for the exertion of oxygen demand during this test is:

$$BOD_t = BODU \cdot (1 - e^{-k_d \cdot t})$$

where BOD_t equals the biochemical oxygen demand observed after t days of incubation, $BODU$ is the ultimate (maximum) BOD, and k_d is the base e , first-order decay rate coefficient.

Each sample was analyzed using two dilutions set up in triplicate. Dilution water samples were developed using two seed volumes and evaluated in triplicate. Samples were incubated for a period of twenty-eight days. The ultimate BOD ($BODU$) and the decay constant, k_d , were determined for each sample using a non-linear, least squares technique.

An acclimated biological culture was developed for each of the different wastes. Samples of a fresh, highly diverse population of microorganisms were taken from the grit chamber effluent of a nearby POTW. The samples were settled and supernate diluted with distilled

water. Substrate requirements were met by sodium acetate while nutrient requirements were based on an assumed protoplasm formulation of $C_{60}H_{87}O_{23}N_{12}P$. Substrate and nutrients were provided daily and the culture was continuously aerated. Acclimation was accomplished slowly by adding increasing volumes of wastewater supplemented with nutrients. As the amount of wastewater was increased, the sodium acetate substrate was reduced until eliminated. By the end of two weeks, the only organic carbon source provided was the particular FTF wastewater. Solids generation and oxygen uptake illustrated microorganism viability.

Organic Content - Surfactants are organic in nature, and thus the chemical oxygen demand (COD) test was used to indicate the total oxidizable carbon in the wastewater. The technique used was the closed reflux, titrimetric method.[2]

Other Waste Parameters - Other measured waste characteristics included pH, alkalinity and total suspended solids (TSS). All were performed in accordance with Standard Methods. [2]

Coagulation Studies

For the waste at hand, removal of the surfactants would all but resolve concerns over foaming and bioinhibition. The literature indicates that surfactants are subject to the same destabilization principles as inorganic sols given they are colloidal emulsions. Therefore, this investigation examined coagulation with trivalent aluminum and trivalent iron for destabilization and removal of AFFF surfactants.

The experimental procedure was basically the same as one would use for an investigation of the coagulation of sols, namely the "jar test."

The coagulants chosen were aluminum sulfate ($\text{Al}_2(\text{SO}_4)_3 \cdot 16\text{H}_2\text{O}$) and ferric chloride (FeCl_3). When pH adjustment was required, a 0.1 N solution of NaOH or a 0.1 N solution of H_2SO_4 was added. A 500-ml sample size was used in conjunction with a Phipps-Byrd six paddle laboratory stirrer.

Samples were rapid mixed on a magnetic stirrer for 2.5 minutes. To provide better mixing and easier measurement of pH, the magnetic stirrer was used rather than the paddle stirrer. Though shorter mixing times are common for water treatment, this mixing period was required to provide a stable pH. After mixing, the sample was immediately placed on the paddle stirrers where it was flocculated for 15 minutes at 30 rpm. Though longer times are typical in water treatment, the 15 minutes interval was chosen to offset effects of the longer mixing time. Samples were settled for 30 minutes after flocculation. Then aliquots of the clarified supernatant were extracted for characterization.

Oxidation Studies

Two oxidants were examined, potassium permanganate and hydrogen peroxide. The permanganate was fed in solution form as was the ferrous ammonium sulfate (FAS) used to supply Fe^{2+} catalyst for the hydrogen peroxide studies. The laboratory procedure consisted of a series of jar tests using the format outlined above. The pH was adjusted immediately after oxidant addition and always within the 2.5 minutes allotted for rapid mixing. After chemical addition and mixing, precipitant flocculation and settling was performed.

RESULTS

To characterize each waste, the ultimate BOD and first-order decay constant, the foaming potential, the COD, the bioinhibitory level, the total suspended solids, the pH, the dilution to eliminate foaming, and the alkalinity were determined. The significance of each was discussed previously. The results of these evaluations are tabulated in Table 1. A brief inspection of this tabulation clearly indicates that the four wastes were very different. This adds credibility to the statement there is no typical waste from a fire training facility. Further, comparison of the synthetic waste's characteristics with those of the grab samples from actual FTFs indicates a marked difference in properties. It seems clear that treatability results derived from the synthetic waste will have little value in defining the treatability of actual wastes.

The synthetic waste was found to have a tremendous foaming potential. Similarly, the dilution required to remove foaming was much greater. These facts indicate that the AFFF concentration of this sample is much greater than should probably be expected for a true FTF wastewater. The BODU and COD of this particular sample was also higher than that of the other samples. Yet, the decay coefficient is significantly lower indicating a poorer biodegradability. These values can also be attributed to a higher-than-normal AFFF concentration.

Table 1
Characterization of Wastewaters from
Firefighter Training Facility

Parameter	Wastewater Sources (*)			
	(1)	(2)	(3)	(4)
Ultimate BOD (mg/l)	138	62	295	3170
First-Order Decay Coefficient, k_d (day^{-1})	0.381	0.533	0.326	0.150
COD (mg/l)	171	174	480	3790
Total Suspended Solids (mg/l)	9	35	28	0
Initial pH	6.9	7.1	6.6	6.1
Alkalinity to pH = 5.10 (mg/l CaCO_3)	65	172	71	4
Bioinhibitory Level (ml/liter)	17	200	40	33
Maximum Foaming Potential (ml per 1000 mls)	310	1375	375	2375
Dilution to Eliminate Foaming (ml diluted to 1 liter)	625	125	187	8

- * - (1) TAFB New Firefighter Training Facility
 (2) TAFB Old Firefighter Training Facility
 (3) CAFB Firefighter Training Facility
 (4) Synthetic Waste

For comparison, the ultimate BOD of AFFF ranges from 300,000 mg/l to 510,000 mg/l. Its COD ranges from 500,000 mg/l to 730,000 mg/l. [27] Based on these values, the mass of AFFF used in synthetic waste development and assuming complete AFFF dissolution, the COD should range between 2,500 and 3,650 mg/l. Similarly, the BODU should fall in a range 1,500 and 2,500 mg/l. Noting complete dissolution did not occur, one can assume that the difference between the actual and hypothetical BODU and COD values can be attributed to the water-miscible fuel hydrocarbons present in the waste.

It was hoped that a correlation could be developed between the bioinhibitory levels and the foaming potential of each waste. After inspecting the data presented in Table 1, it appears there is no correlation between these two parameters. This may be due to different concentrations of water-miscible fuel hydrocarbons present in the different wastewaters. Also, the presence of unidentified components could have an impact.

The low alkalinity of the synthetic waste can be attributed to distilled water being used for its development. Similarly, the water used at FTFs would be the primary source of this parameter. The pH values are consistent with alkalinity.

The total suspended solids content of the wastewaters from CAFB and TAFB's old FTF were appreciable while the synthetic waste and the TAFB's new FTF wastewater were almost devoid of TSS. Observations at each facility confirmed this with water in the TAFB's old FTF's oil/water separator laden with bacteria and algae. The water in the pond at CAFB's FTF had suspended clay particles which apparently

resulted from the new liner placed on the pit area. The low level of solids found for the new TAFB facility might be attributed to several factors. One would be that the oil/water separator has about four times the surface area of CAFB's FTF and is a newer design than that of TAFB's old FTF, resulting in better solids separation. Another factor might be that the new TAFB FTF had been used for only a few months prior to sample collection, providing insufficient time for an algae population to develop.

This brings up another reason for the lack of correlation between foaming and bioinhibition. If the surfactants in AFFF were to adsorb to the sooty particulate matter present in some of the wastes, its effectiveness at generating foam would be impaired. At the same time, microbial organisms would be attracted to these same sites, thereby coming nearer the surfactant materials. If this is true, the foam test alone will never be a good indication of the surfactant concentration present in, or the inhibitory strength of, a waste.

Chemical Coagulation

Initial evaluation of alum coagulation used the synthetic wastewater. Due to its low alkalinity, some pH adjustment to achieve optimum treatment was anticipated. However, to obtain an initial indication of optimum coagulant dose, the first test was performed without pH modification. This test found that foaming potential was eliminated at dosages above 40 mg/l as Al^{3+} . It was noted that at this and greater alum dosages no sludge was produced though the wastewater turned milky white due to pin floc development. It was also noted that the pH fell

below 5.0 units at or above 40 mg/l as Al^{3+} which is outside the optimum insolubility range of aluminum hydroxide.

Additional testing with alum was performed but with the final pH of the reaction solution being adjusted to various levels with 0.1 N NaOH. According to the literature, a final pH of 5.7 units would ensure optimum coagulation and better sludge production. Table 2 summarizes results from the tests performed. From this data one notes that, at pH levels above 5.0 units, the foaminess of the wastewater did not vary with alum dose. During these experiments sludge production was never observed though a fine pin floc did consistently form.

From this one can conclude that coagulative-based removal of the AFFF surfactant was not providing the foaminess reduction observed previously. However, the values in Table 2 would indicated that pH adjustment alone was insufficient to produce this effect. This was confirmed by additional tests with sample pH being adjusted between 3.0 and 11.0 units and no coagulant addition.

Throughout the performance of these coagulation studies the floc generated appeared light and fluffy with poor settling characteristics. Referred to as pin floc, it can result from long mixing times if energy is added to the point of destroying developed particles. It is also a good indication of improper pH.

Coagulation of the synthetic waste with ferric chloride as the coagulant was also attempted. Again, the foam inhibition effect observed in the alum studies was encountered. In particular, dosages above 50 mg/l as Fe^{3+} cut out foaminess. However, foam was not

Table 2

Results for Coagulation with
Aluminum Sulfate of Synthetic Waste

Dosage Alum (mg/l Al^{+3})	Volume of Foam Generated (ml/l) at stated final pH			
	Not Adjusted	4.9	5.7	5.3
0	2375	2375	2375	2375
20	2000	1875	2375	2375
25	*	*	2375	2375
30	*	1750	2375	2375
35	*	*	2375	*
40	750	450	2375	2375
45	*	*	2375	*
50	*	0	*	2375
60	0	0	2375	*
80	0	0	*	2375
100	0	0	2375	*

* - Value not reported due to data collection error or analysis not being performed.

inhibited until the pH fell below 4.4. Further, pin floc did not develop during these jar tests.

Because sludge development did not occur, it was anticipated that the organic content or biodegradability of this waste would not change. The results of the COD and BOD tests confirmed this. The data from the bioinhibitory test similarly indicated no significant change in this parameter.

At this point testing of the samples collected from the three FTFs began. In each case, the wastewater was tested initially with varying dosages of aluminum sulfate, without pH adjustment. A summary of these initial tests is presented as Table 3. From inspection of this tabulation one can see that the pH-foaminess relationship existed when alum was added. Foam production was nonexistent when pH reached the 5.0 level. Tests involving pH adjustment with and without coagulant were performed on each sample. The results confirmed that the phenomena observed previously for the synthetic waste exists for the three FTF wastes as well. However, there were some reaction differences which should be noted.

When the CAFB wastewater sample was treated with alum dosages of between 40 and 80 mg/l as Al^{3+} , good floc formation was observed and removal of solids was achieved. Noting the increase in foaminess which occurred for this level of treatment, it would appear that the suspended solids in the wastewater broke the surface tension created by the surfactants. This would greatly inhibited the true foaming potential of the CAFB waste stream. By destabilization and removing these solids the surface tension decreased, allowing an associated increase in foam

Table 3

Foam Generation Results from the
Aluminum Sulfate Coagulation Studies
of Firefighter Training Facility WastewaterS

Dose Alum (mg/l Al ⁺³)	CAFB's FTF		TAFB's Old FTF		TAFB's New FTF	
	Foam Produced (ml/l)	Final pH	Foam Produced (ml/l)	Final pH	Foam Produced (ml/l)	Final pH
0	375	6.5	1375	7.0	310	6.9
20	438	6.5	*	*	*	*
40	3875	6.2	*	*	*	*
50	1000	6.2	*	*	375	6.8
60	1375	6.2	*	*	*	*
80	1375	6.0	*	*	*	*
100	1125	5.9	1500	6.82	438	6.5
125	875	5.9	*	*	*	*
150	375	5.5	*	*	125	5.4
160	250	5.4	*	*	*	*
170	125	5.3	*	*	*	*
175	0	5.2	*	*	*	*
200	0	4.9	1500	6.4	0	4.7
300	0	4.6	1500	6.3	*	*
350	*	*	750	6.1	*	*
400	0	4.5	500	5.8	*	*
415	*	*	0	5.1	*	*
450	*	*	0	4.8	*	*

* - Value not reported due to data collection error or analysis not performed.

generation. This continues until the pH inhibiting effect begins to take over.

With regard to the organics present, COD was reduced by about 25 percent at a dosage of 80 mg/l as Al^{3+} . There was no decrease in the ultimate BOD and the decay rate remained effectively the same. Bioinhibition characteristics of this waste similarly remained consistent. It would therefore appear that the solids removed are chemically oxidizable but not readily degradable by microorganism.

The total suspended solids concentration of the wastewater generated by the older FTF at TAFB was similar to the CAFB waste. It appears that there was sufficient alkalinity for the aluminum precipitation reaction. When tested, only a minimal increase in the foam production was realized with the removal of the suspended matter. The higher initial alkalinity retarded pH depression and required a large alum dosage to eliminate the foaming potential. Again, this point was reached at a pH of about 5 units.

For the waste from the old FTF at TAFB, COD was reduced by about 25 percent at what appeared to be an optimum dose of 200 mg/l as Al^{3+} . However, the BODU of this sample was reduced by fifty percent and the decay coefficient was reduced to 0.36 days^{-1} . This can be attributed to the solids contained in this waste being microorganisms rather than the clayish material found in the CAFB waste.

The wastewater generated at the new training facility at TAFB was characteristically different from the others. When the pH was reduced to 4.7 with an alum dose of 200 mg/l as Al^{3+} , a light non-settling pin floc developed similar to that observed during the tests with the

synthetic waste. Other similar results included foam production stopping with BOD and COD levels effectively unchanging at this alum dose.

A method that may be applied to enhance coagulation is the addition of a weighting agent. The coagulant aid chosen was kaoline, an adsorptive clay which is readily available and often used in the production of potable waters. A series of jar tests were set up to see if addition of the clay would produce surfactant removal. Samples of synthetic waste and waste from TAFB's new FTF were used. The result was that a very good floc formed which provided for good solids separation. However, foaminess was not reduced nor was the COD of the treated waters.

Coagulation of the FTF wastewaters with ferric chloride proved fruitless. As with the synthetic waste, using this coagulant necessitated lower pH levels than was needed with alum to end foaming. Solids separation results were similar to those of alum with good removal for the wastewaters from the CAFB and TAFB's older facilities.

Chemical Oxidation

The initial valuation of hydrogen peroxide as a chemical oxidant was performed on the synthetic waste. At the outset, H_2O_2 dosages ranging from 0.5 to 50 mg/l were examined without pH adjustment or Fe^{2+} catalyst addition. These tests were to determine if a threshold or an optimum dose existed within this range. Unfortunately, a measurable change in waste characteristics did not take place over this range of H_2O_2 addition. Initially it was thought that greater dosages might be

required to offset an unknown competing reaction. However, titration with ferrous ammonium sulfate indicated that an oxidant residual existed at the lowest dose level after one hour. It was therefore surmised that oxidation with hydrogen peroxide was not possible under natural conditions.

The literature indicates that an acidic environment is beneficial for peroxide oxidation of alkanes. To test if pH was a factor, samples of the synthetic waste were dosed with 10 mg/l of H_2O_2 and the pH was varied from 3.0 and 11.0 units. Once again, no significant change in foaminess, COD, or BODU were noted.

The last test of hydrogen peroxide oxidation of the synthetic waste examined Fe^{2+} catalyst addition. A dose of 10 mg/l of H_2O_2 was added to several samples. A pH of 4.0 units was used for the reaction, as based on the literature. This adjustment was made immediately after hydrogen peroxide addition. Addition of FAS solution followed pH adjustment. Fe^{2+} catalyst concentrations up to 50 mg/l were examined.

The results of this experiment were generally unimpressive. At catalyst levels above 35 mg/l, final waste characteristics did not change significantly. It was noted that sedimentation of an iron precipitant became significant for the samples. Below 35 mg Fe^{2+} /l a reduction in COD was observed. A catalyst dose of 12 mg Fe^{2+} /l yielded the greatest COD removal. For this dosage, COD was reduced by 24.2 percent, BOD reduction was 20.0 percent and the decay coefficient remained about the same at 0.16 day^{-1} . However, foaminess was effectively unchanged over the entire range of FAS additions.

For the three FTF samples collected for this project, somewhat limited testing of hydrogen peroxide oxidation was performed. A solution pH of 4.0 units was used. A catalyst to hydrogen peroxide ratio of 1.2 was used for all samples and was based on the optimum value for the synthetic waste. The tests evaluated H_2O_2 dosages ranging from 1 to 10 mg/l. Once again, none of the samples showed any reduction in foam generation after treatment. COD removal efficiencies were low. For these reasons, BOD was not performed. These results would seem to eliminate hydrogen peroxide from consideration as an effective oxidizing agent for the pretreatment of FTF wastewaters.

The synthetic wastewater was the first sample examined to evaluate the potential usefulness of potassium permanganate in pretreatment. These studies examined the addition of $KMnO_4$ up to a level of 90 mg/l. The data showed that COD removal of 21.8 percent was achieved at the 40 mg/l $KMnO_4$ level and that doses above this did not significantly reduce the level of this parameter further. Below the 40 mg/l $KMnO_4$ level, the BOD and the decay coefficient were reduced. Above this, residual $KMnO_4$ inhibited bacterial activity.

It therefore seems that the potassium permanganate reacted with some of oxidizable organics which were also biodegradable. However, once these materials were stabilized, it appears that the reaction kinetics changed, slowing the rate of oxidation significantly. Further, foam generation for all of the samples was unchanged. It would therefore appear that the surfactants were not involved in the more rapid oxidation reaction.

The CAFB sample was exposed to a similar set of tests though there was a significant difference in results. The data from these experiments are presented Table 4. In short, potassium permanganate doses of 20 mg/l and greater resulting in a reduced BODU concentration. For a measurable COD reduction, 30 mg/l was required. It is also interesting to note an increase in apparent biodegradability, as indicated by the decay coefficient, at the 10 mg/l dose level. No foam was developed from any of the treated samples.

Table 4
Potassium Permanganate Oxidation of CAFB Wastewater

Chemical Dose (mg/l KMnO_4)	COD (mg/l)	Ultimate BOD (mg/l)	Decay Coefficient (day^{-1})	Foam Produced (ml/l)
0	537	313	0.316	0
10	531	310	0.377	0
20	535	268	0.255	0
30	445	122	0.158	0
40	484	129	0.131	0

This is the desired result from pretreatment of this waste before biological treatment in a POTW. However, more than oxidant addition must have played a role. At the 10 mg/l level the dose/COD ratio for this waste was 0.019 mg/mg. At the same time the synthetic waste had been exposed to a 0.024 mg dose/mg COD and had not produced these results.

Reevaluation of the data indicated that the final pH of the synthetic waste samples ranged between 6.8 for the lowest dose to 7.7 at the highest dose. The CAFB samples had final pHs of about 6.6. Prior to this discovery, pH adjustment was not considered because of the desirably neutral level of the raw waters and the intent to only pre-treat. However, a wastewater having a pH of 6.5 would be acceptable for discharge into a POTW under current USEPA guidelines and the literature does indicate that acidic conditons enhance the permanganate oxidation process. Therefore, the synthetic waste was retested using a final pH of 6.5. Under these conditions, a level of KMnO_4 addition was observed which eliminated the foaming potential of the synthetic waste.

It appears from these experiments that a 40 mg/l of KMnO_4 is sufficient to provide for the destruction of the foam generation potential of the synthetic wastewater. It should be noted that at the permanganate dosage which abated foam generation, COD had been reduced only by about six percent and that there was virtually no removal of BOD. Further, the decay coefficient had increased only slightly at this treatment level. Table 5 contains the values of these parameters at the dose of KMnO_4 which eliminated foaminess. The values for foam generation potential as determined during these experiments is presented in as Figure 2.

Given these results, the samples collected from the two TAFB facilities and from CAFB were evaluated for potassium permanganate oxidation at a pH of 6.5 units. Figures 3 through 5 illustrate the test results on foaming. Using these data, optimum dosages have been defined and are summarized in Table 5. The waste characterization for these

Table 5

Results of Potassium Permanganate Addition to Wastewaters
from Firefighter Training Facilities (Final pH = 6.5)

Parameter	Wastewater Sources (*)			
	(1)	(2)	(3)	(4)
KMnO ₄ Dose Required to Eliminated Foaming (mg/l)	4	10	5	40
COD (mg/l); Initial	164	170	465	3660
Final	158	156	455	3440
Ultimate BOD (mg/l); Initial	133	53	270	3090
Final	127	58	250	3060
Decay Coefficient, k_d (day ⁻¹);				
Initial	0.38	0.53	0.33	0.15
Final	0.39	0.48	0.32	0.15

- * - (1) TAFB New Firefighter Training Facility
 (2) TAFB Old Firefighter Training Facility
 (3) CAFB Firefighter Training Facility
 (4) Synthetic Waste

Figure 2

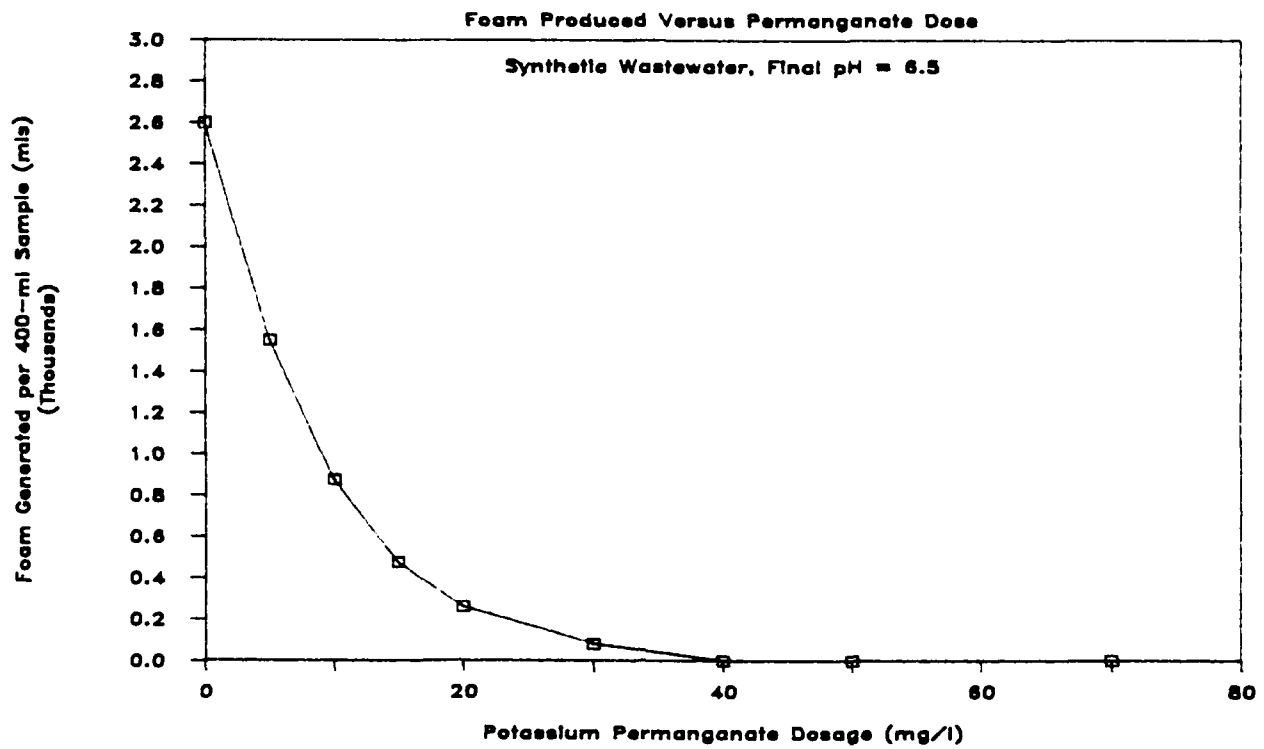


Figure 3

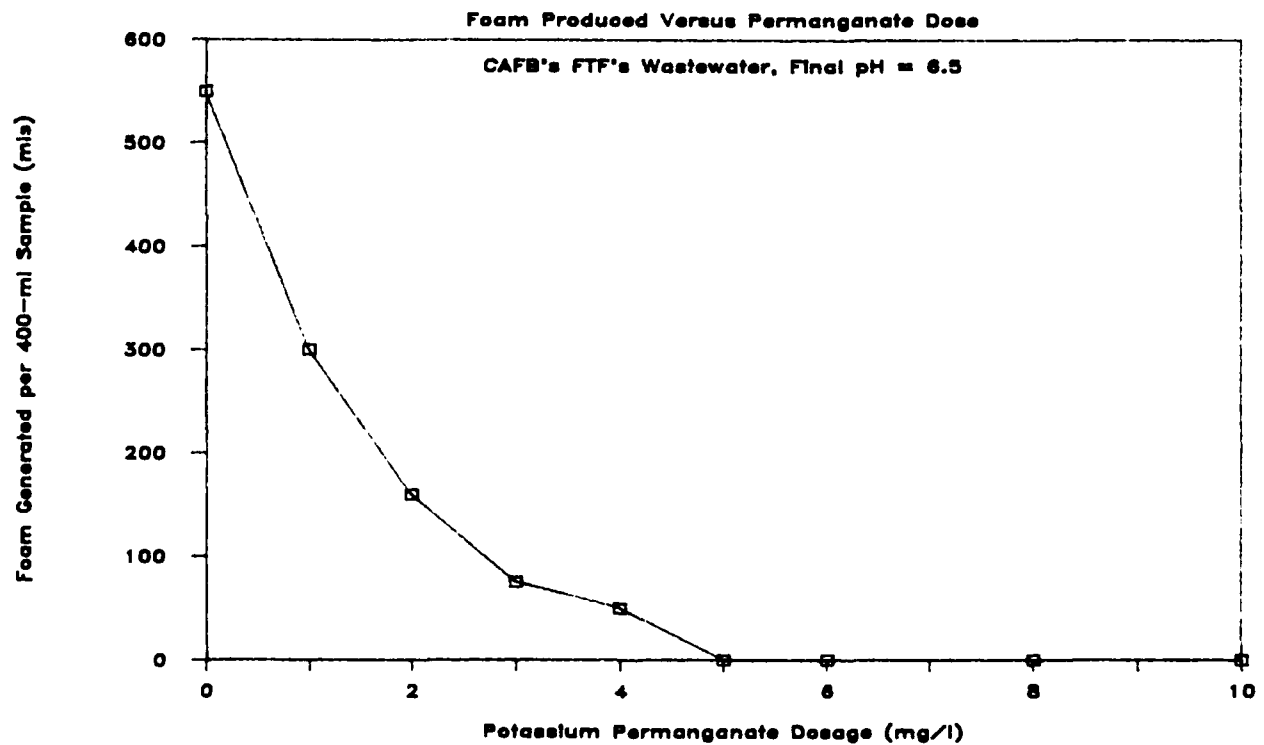


Figure 4

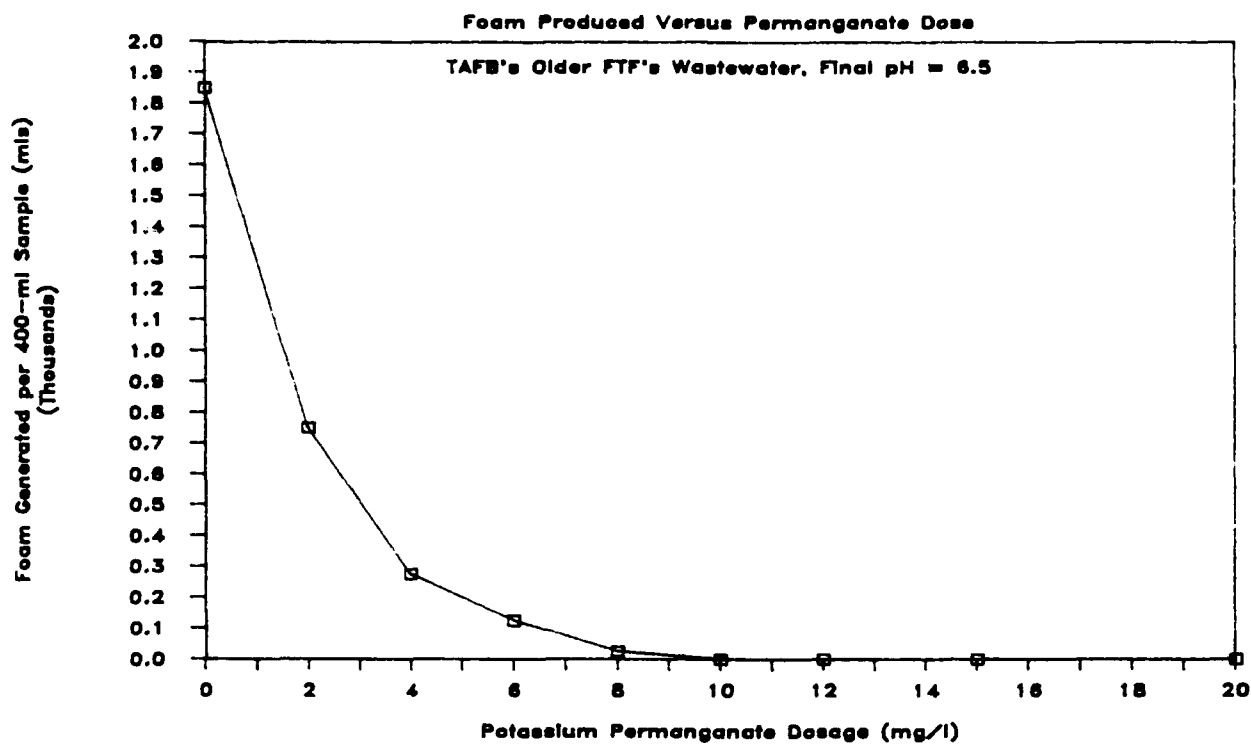
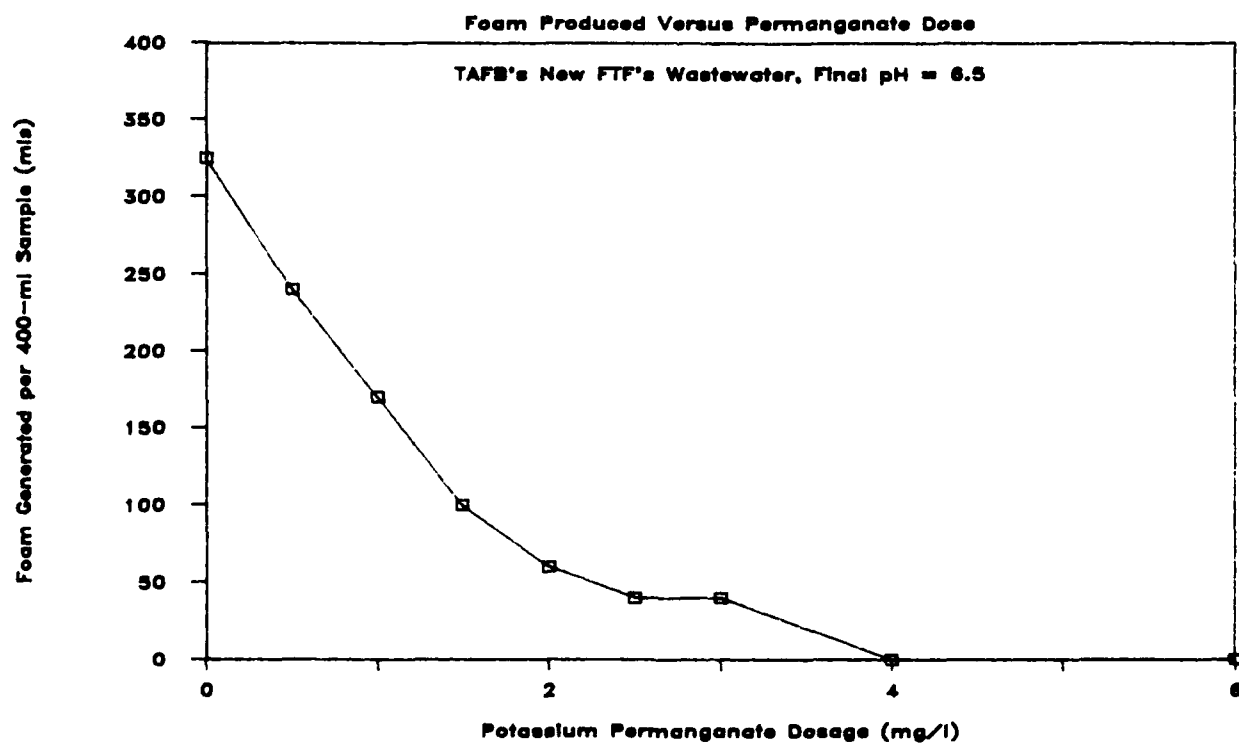


Figure 5



levels of potassium permanganate addition is also presented in this tabulation.

From these experiments, it has been found that potassium permanganate is effective at destroying the foaming potential of FTF wastewaters if the pH of the solution is 6.5 units. At the same time, there does not seem to be a significant shift in organic content (COD) or biodegradability (Ultimate BOD and decay rate).

CONCLUSIONS AND RECOMMENDATIONS

The experiments presented in this report examined the possibility of reducing the foaming potential firefighter training facility wastewaters. The improvement in its biodegradability was considered a desirable side effect.

Potassium permanganate addition coupled with pH control has been found successful at reducing the foaming potential of wastewaters generated by FTFs. No significant sludge was produced, foaming was permanently eliminated, and biodegradability may actually be enhanced. The variability of the characteristics of these wastes impacts pretreatment needs and chemical requirements. Therefore, studies will have to be performed on a case-by-case basis before wastewater discharge to a POTW.

Coagulation for the removal of AFFF surfactants was found to have little if any application in pretreating these wastewaters. A foam inhibition effect was found early in the experimental process. This occurred when pH of the solution fell into the range of 4.8 to 5.1 upon the addition of alum. However, adjusting the pH out of this range, as would be required before waste discharge to a POTW, resulted in the foaming potential of the waste being reinstated.

The samples collected from the CAFB and older TAFB facilities contained solids which were readily coagulated, flocculated and removed through sedimentation. This validated the mixing, flocculation, and sedimentation procedures used during this investigation. During

coagulant addition, foaming potential of these samples increased to a point of inflection after which the foam production decreased to zero. It appears that the suspended solids in the water broke the surface tension of the bubble created by the AFFF surfactants. With removal of this matter, the surface tension and foam production increased. The inflection point of the curve occurs when the pH reaches the level for foaming inhibition.

The synthetic sample and the wastewater from TAFB's new FTF contained no significant solids and coagulation was unable to produce a floc particle that would settle. Kaoline was added as a weighting agent and particle sedimentation did occur. This addition did not improve the post-treatment characteristics of the wastewaters. It was observed that with the kaoline in suspension the foaminess of the waste was reduced. However, once solids separation had been achieved, the foaming potential returned to its original level.

The BOD data for the samples treated with alum and ferric chloride showed no increase in the biodegradability of the wastes after treatment. This would be expected given that no solids of significance were removed.

Studies identified the level of waste dilution at which foaming potential was eliminated. This data appeared to be first-order in function and an effort to regress an equation relating initial foaming potential to dilution was made. Unfortunately, the confidence contours were so large that no distinction could be made between the functions. This can be attributed to a limited data base.

It is known that the biological inhibition is dependent on the surfactant concentration present in the wastes. The level of wastewater dilution needed to eliminate foam production should have correlated with the dilution below which bioactivity was unimpeded. The data did not correlate in this manner. This can be attributed, in part, to the fact that foam generation is sensitive to suspended solids for any given concentration of AFFF.

Based on the results of this investigation, several questions seem to require further study. Some of these are listed below.

1. What is the affect pH has on permanganate oxidation of this waste?
2. What impact, permanent or temporary, do inert and biological solids have on the foaming potential of FTF wastes?
3. Can peroxide oxidation be enhance with UV or will the spectral ban of the surfactants impact this process?
4. Can a relationship between solids, foaming potential, and/or dilution required to eliminated foaming be developed which predicts surfactant concentration? Predicts bioinhibitory concentration?

REFERENCES

1. "Action Description Memorandum for Live Fire Training Facility, Tyndall AFB, FL", Prepared by Oak Ridge National Laboratory, Martin Marietta Energy Systems, Inc. for AFESC under IAG 40-1762-86, 1986.
2. Zachritz, W. H., Jr., Evaluation of Four Systems to Treat Fire Training Pit Wastewaters, AFESC Completion Report, Tyndall AFB, FL, 1987.
3. Chan, D. B., "Disposal of Wastewater Containing Aqueous Film Forming Foam (AFFF)", U. S. Navy Civil Engineering Laboratory, Port Hueneme, California, Tech Memorandum M-54-78-06, 1978.
4. Carlson, R. E., "The Biological Degradation of Spilled Jet Fuels: A Literature Review," USAF/AFESC Tech Report ESL-TR-81-50, 1981.
5. APHA, AWWA, WPCF, Standard Methods for the Examination of Water and Wastewater, 16th ed., Am. Public Health Assoc., Washington DC, 1985.
6. Truax, D. D., The Roles of Biodegradation and Adsorption in the Biological Activated Carbon Reactor, A Ph.D. Dissertation, Mississippi State University, 1986.
7. Schulz, W. D., "Characterization of Fire Training Facility Wastewater," USAF-UES Summer Faculty Research Program Final Report, Contract No. F49620-85-C-0013, 1987.
8. Telephonic communication with Mr. Bruce Nielsen of AFESC/RDVW on October 15, 1987.
9. Truax, D. D., "Ozonation of Firefighter Training Facility Wastewater and Its Effect on Biodegradation," USAF-UES Summer Faculty Research Program Final Report, Contract No. F49620-85-C-0013, 1987.
10. "Physical-Chemical Treatment of Wastewater from Navy Firefighter Schools," Prepared for the Naval Facilities Engineering Command by Engineering Science, Inc., Contract No. N00025-74-C-0004, 1976.
11. Federal Guidelines: State and Local Pretreatment Programs - Appendix 8; Volume III, EPA Construction Grants Program Information, EPA-430/9-76-017c, USEPA, Washington, D.C., 1977.
12. In-Plant Control of Pollution: Upgrading Textile Operations to Reduce Pollution, USEPA Tech. Transfer, EPA-625/3-74-004, Cincinnati, OH, 1974.

13. Nemerow, N. L., Liquid Waste of Industry: Theories, Practices, and Treatment, Addison-Wesley Pub. Co., Reading, MA, 1971.
14. Bikerman, J.J., Foams, Theory and Industrial Applications, Reinhold Publishing Corp., New York, 1953.
15. Rosen, M.J., Surfactants and Interfacial Phenomena, John Wiley & Sons Inc., New York, 1978.
16. Mysels, K.J., Introduction to Colloid Chemistry, 2nd ed., Interscience Publishers, New York, 1978.
17. Becher, P., Emulsions: Theory and Practice, 2nd ed., Reinhold Publishing Corp., New York, 1965.
13. Sherman, P., Emulsion Science, Academic Press, New York, 1969.
19. Weber, W. J., Physicochemical Processes for Water Quality Control, John Wiley & Sons, Inc., New York, 1972.
20. Stumm, W. and O'Melia, C.R., "Stoichiometry of coagulation", J.AWWA, 60, 514-539, 1968.
21. Hummel, Dieter, Identification and Analysis of Surface Active Agents by Infrared and Chemical Methods, John Wiley & Sons, New York, 1962.
22. Ladbury, J.W. and Cullis, C.F., "Kinetics and Mechanism of Oxidation by Permanganate", Chemical Reviews, Vol. 58, Williams and Wilkins Co., Baltimore, MD, 1958.
23. Haines, A. H., Methods for the Oxidation of Organic Compounds: Alkanes, Alkenes, Alkynes, and Arenes, Academic Press, New York, 1985.
24. Solomons, T.W.G., Organic Chemistry, 3rd ed., John Wiley & Sons, New York, 1984.
25. Merz, J.H. and Waters, W.A., "Some Oxidations involving the Free Hydroxyl Radical", J. of Chemical Society, London, S15, p. 2427, 1949.
26. Sims, A.F.E., "Phenol Oxidation with Hydrogen Peroxide", Effluent and Water Treatment Journal, V21, n3, pp. 109-112, 1981.
27. Ventullo, R. M. "Biodegradation of Aqueous Film Forming Foam Components in Laboratory Scale Microcosms", A Report Submitted to Universal Energy Systems, Dayton OH, 1987.
28. Telephonic communication with Mr. Bruce Nielsen of AFESC/RDV on April 19, 1988.

**STRESS TRANSMISSION AND MICROSTRUCTURE IN
COMPACTED MOIST SAND**

by

George E. Veyera, Ph.D.
Assistant Professor of Civil Engineering
and
Blaise J. Fitzpatrick, Graduate Student

Department of Civil and Environmental Engineering
University of Rhode Island
Kingston, RI 02881

FINAL REPORT

for

RESEARCH INITIATION GRANT PROGRAM

Contract No.: F49620-88-C-0053/SB5881-0378

Sponsored by

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

31 December 90

STRESS TRANSMISSION AND MICROSTRUCTURE IN COMPACTED MOIST SAND

by

George E. Veyera, Ph.D.
Assistant Professor of Civil Engineering
and
Blaise J. Fitzpatrick, Graduate Student

Department of Civil and Environmental Engineering
University of Rhode Island
Kingston, RI 02881

ABSTRACT

The research described in this report was conducted to examine the relationship between compaction moisture content, dynamic stress transmission and soil microstructure in compacted unsaturated Ottawa 20-30 sand. Recent research has shown that moisture conditions during compaction can increase the stress transmission ratio measured in dynamic impact tests by as much as a factor of two and can also lead to increased stress wave propagation velocities. Other studies have demonstrated that both the method of compaction and the amount of moisture present during compaction have a measurable influence on both the static and dynamic properties of sands. Experimental evidence suggests such behavior can be attributed to variations in soil microstructure and compressibility as a result of conditions during compaction. However, a clear and concise explanation of the phenomenon and the interrelationship among various parameters is not currently available.

Split-Hopkinson Pressure (SHPB) tests were conducted on compacted moist specimens of Ottawa 20-30 sand tested moist and after oven drying. The SHPB results indicate that the moisture present during compaction has a significant effect on stress transmission even for specimens compacted moist and tested dry. These results also

suggest that moisture is an important factor contributing to the development of soil microstructure and soil stiffness during compaction. In addition, compacted specimens were epoxied and sectioned for microstructural analysis using an approach based on standard petrographic procedures available in the open literature. The preliminary results obtained to date indicate that some preferred particle orientations do exist with variations in moisture conditions at compaction which could affect soil behavior. However, no strong preferential orientation has been observed in the data so far. Further detailed microstructural studies are currently being conducted and will be the subject of future reports and publications by the authors.

ACKNOWLEDGEMENTS

We would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research and the Air Force Engineering and Services Center for sponsorship of this research. The College of Engineering and the Department of Civil and Environmental Engineering at the University of Rhode Island have also provided support for this research. Universal Energy Systems provided assistance in all administrative and directional aspects of this program. We would like to especially thank Dr. C. Allen Ross for his interest and enthusiasm in the research topic. His continual support, encouragement and keen insight are very much appreciated. Special thanks are also due to Captain Steven T. Kuennen for his assistance in performing tests and in supporting the research. The technical library at AFESC has been invaluable in obtaining numerous reference items from a variety of sources, some of which were often difficult to locate. The experimental laboratory portion of our work was conducted during the summer months at Tyndall Air Force Base and we would also like to express our very sincere thanks to Mr. L. Michael Womack at RDCM for his hospitality, support and interest in this research. Finally, we would like to thank all of the many staff members at HQ AFESC and at RDCM for their camaraderie and friendship during the summer portion of this research and for making our stay at Tyndall AFB so productive and enjoyable.

I. INTRODUCTION

The prediction of ground motions from explosive detonations and their effects on structures requires information about the response of geologic materials to intense transient loadings. Both laboratory and field investigations have provided insight into the stress wave propagation characteristics of soils. Of particular interest is the ability of a soil to transmit applied dynamic stresses. Since soil is a multiphase media, in the general case (eg. solids, water and air), its static and dynamic behavior is very complex. The general nature of stress wave propagation in particulate media such as soils depends on a number of parameters, the interrelationship between which is not fully understood. Soil microstructure is affected by the degree of saturation (moisture condition), soil density, effective stress, applied stress intensity, stress history and the nature of the material itself (e.g. particle size, shape, distribution, mineralogical constituent(s), etc.).

The ability of a soil to transmit applied dynamic stresses (energy) is of particular interest to the U.S. Air Force with respect to military protective construction and survivability designs. Typical engineering analyses assume that little or no material property changes occur under dynamic loadings, and in addition, analyses do not account for the effects of saturation (moisture conditions) on energy transmission in soils. This is primarily due to an incomplete understanding of the behavior of soils under transient loadings and uncertainties about field boundary conditions.

Current analysis and prediction methods generally use material properties data based on conventional weapons detonations in dry, or to a limited extent, saturated soils. Results from U.S Air Force field and laboratory tests with explosive detonations in soils have shown that material property changes do in fact occur and that variations in soil stiffness (or compressibility) significantly affect both dynamic and static stress transmission. However, there are currently no theoretical, empirical or numerical methods available for predicting large amplitude compressive stress wave velocity and stress transmission in unsaturated soils (Crawford et al., 1974; Ross et al., 1986; WES, 1984).

Dynamic stress transmission has been shown to be dependent upon the moisture content and boundary conditions present during compaction (Ross et al. 1986; Ross 1989; Veyera 1989). It is believed that the observed behavior can be attributed to variations in soil microstructure developed as a consequence of moisture present during compaction. Considering this, the research outlined herein was performed as a part of the 1990 Research Initiation Program (RIP) to obtain information on the development of microstructure in compacted unsaturated sand and its relationship to stress transmission behavior. A series of tests were conducted to assess the stress transmission characteristics of compacted moist Ottawa 20-30 sand. In addition, a detailed examination of the microstructure formed during compaction was initiated. Specific microstructural parameters of interest include particle orientations, void space orientation, and percentage void space as a function of moisture content (saturation) and confinement conditions and in particular, the spatial variations occurring on vertical and horizontal planes. Procedures available in the open literature for studying planar pore and grain patterns have been adopted as a basis for analyzing the data obtained.

The results of studies such as this will lead to a better fundamental understanding of the role of microstructure as it affects the macroscopic engineering behavior of soils. The microstructural characterization of unsaturated soils will be a key element in establishing and developing an understanding of stress transmission in unsaturated soils and will have direct applications to groundshock prediction techniques including stress transmission to structures.

II. RESEARCH OBJECTIVES

The study described herein was conducted to examine the development of microstructure in compacted moist sand and its influence on dynamic stress transmission. The primary objectives of the RIP research study were as follows:

1. To determine the influence of moisture content during compaction on soil microstructure and soil properties (compressibility, wave speed, transmission ratio, strength); and
2. To investigate and qualitatively describe the development of microstructure in compacted unsaturated sand and its effect on stress transmission from conventional weapons effects.

III. BACKGROUND

A. Stress Transmission in Unsaturated Soils

Recent experimental investigations by Ross et al. (1986) and Ross (1989) have demonstrated that compacting moist sands at varying degrees of saturation prior to dynamic testing can increase the stress transmission ratio by as much as a factor of two and can also lead to increased stress wave propagation velocities. Additional research by Charlie and Pierce (1988) to study the influences of capillary stresses on the behavior of the same soils tested by Ross further confirmed and extended those findings. They also demonstrated that the influence of capillary pressures on the stiffness of sands is negligible, being on the order of about 7 kPa (1 psi) or less which is insignificant in comparison with high intensity transient dynamic loadings. Therefore, it appears that capillary pressures do not directly affect a granular soil's ability to transmit stresses. However, studies performed by Ross et al. (1986), Ross (1989), Veyera (1989) and as a part of this investigation suggest that capillary pressures may strongly influence the soil microstructure developed during compaction (including soil placement and soil formation in the field) which could significantly affect both the static and dynamic behavior of soil. Recent studies by Charlie

and Walsh (1990) were conducted to determine the response of unsaturated soils to small scale explosive detonations in the centrifuge. Their results have also shown stress transmission to depend on compaction moisture conditions.

Studies by a number of other investigators have also shown that the compaction of sands with moisture present has a measurable influence on both static and dynamic soil properties which can be attributed to variations in soil stiffness and microstructure. Mitchell et al. (1976) studied the effects of sand fabric and sample preparation method on the liquefaction behavior of sands. They performed cyclic triaxial tests and observed significant differences in behavior based on how specimens were prepared. Mulillis et al. (1977) investigated 11 different packing methods and showed that the method of compaction, particularly the initial moisture conditions, strongly influences the cyclic liquefaction behavior of fine sands. Wu et al. (1984) performed resonant column tests on fine sands and observed that capillary pressures in specimens compacted moist at saturations in the range of from 5% to 20% produced a significant increase in the dynamic shearing modulus.

An interesting study to investigate the effects of saturation on wave velocity in sandstones was reported by Hughes and Kelly (1952) who directly measured variations in dilatational velocity with saturation on rock core specimens using pulse techniques. As a part of the study, both temperature and confining pressure were varied. Results indicated that wave velocity increased between 0% and about 20% saturation, remained essentially constant between about 20% and 90% saturation, and then decreased thereafter for pressures between 105 and 526 kPa (725 and 3626 psi) at room temperature (see Figure 1). At higher pressures, the wave velocity was fairly constant up to about 90% saturation. Wyllie et al. (1956) performed tests similar to those of Hughes and Kelly (1952) on Berea sandstone and further substantiated the previous findings. These results are remarkably similar to those obtained by Ross et al. (1986), Ross (1989), Veyera (1989) and the study

Core No. 188, Sample No. 46 - Silty Shale

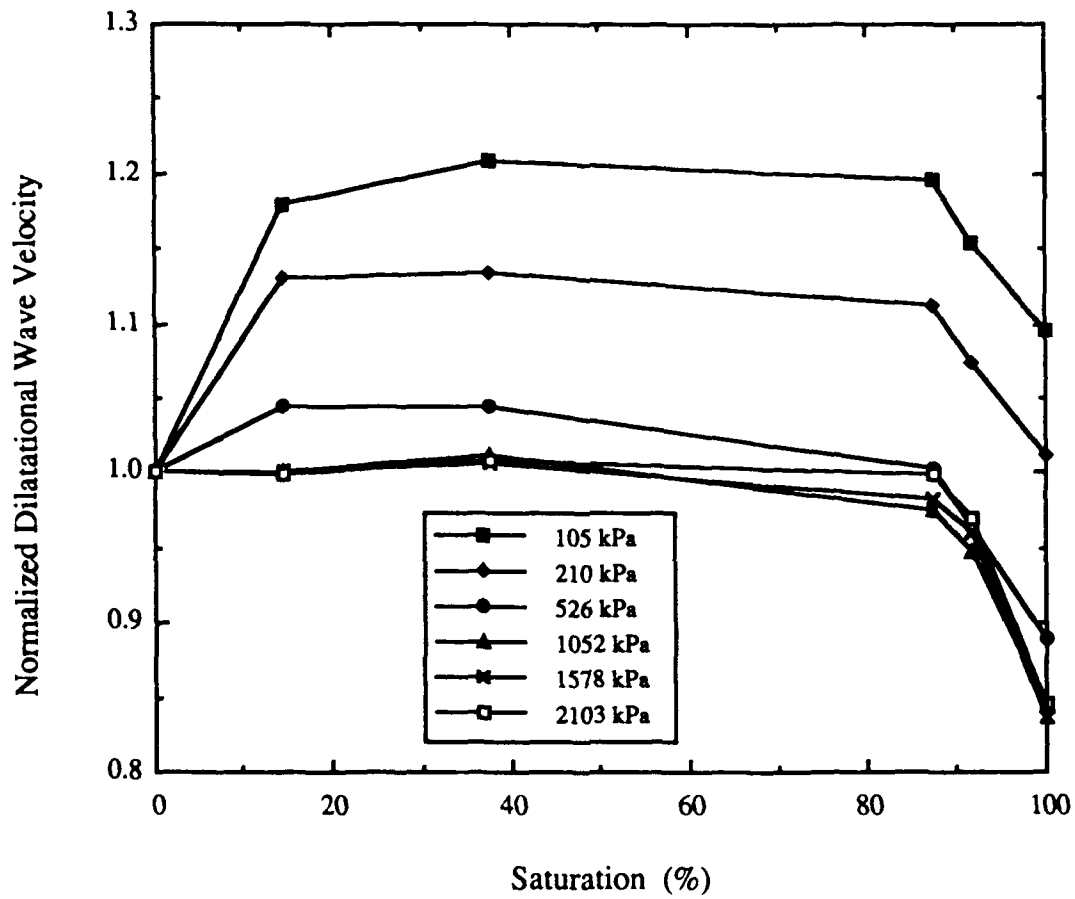


Figure 1. Normalized Dilatational Wave Propagation Velocity as a Function of Saturation and Confining Pressure (Hughes and Kelly, 1952).

presented in this report for SHPB tests performed on uncemented sands compacted at different saturations and a constant dry density.

B. Microstructural Analysis

Various studies have been conducted in an attempt to characterize and relate soil structure at the microscopic level to the engineering properties of soils (macroscopic level). These investigations have resulted in the development of techniques for microstructural analysis of soil. The procedures generally involve thin sectioning of epoxied specimens and the characterization of directional grain and pore space orientations by statistical analysis. Then a correlation with engineering behavior can be extrapolated from the results. Standard petrographic analysis procedures, commonly used in geologic studies, are available in the open literature (Brewer, 1962; Turner and Weiss, 1963). These techniques have been modified for use in geotechnical engineering analyses of soil microstructure (Campbell, 1985; LaFeber 1965, 1972; Oda 1972a, 1972b) and have been adopted for use in this study. A few particular approaches and observations are presented in this section.

LaFeber (1965) developed a method for analyzing soil microstructure in terms of the spatial orientation of planar pore patterns. He indicates their significance by stating that these patterns "...can be expected to be an expression of the depositional and/or stress-strain history of the particular soils. As such, they should be of paramount importance in the study of the mutual relations between soil fabric and soil mechanical behavior." LaFeber (1972) further extended his work to the analysis of three-dimensional grain orientations and anisotropic fabrics by studying compound linear and planar fabric patterns and demonstrated the influence on soil fabric on the mechanical properties of several soils. Oda (1972a, 1972b) noted that particle arrangements in sands are determined by both the particle shape and the method of compaction and that initial fabric is important to mechanical properties. He also developed a method for characterizing fabric features in

terms of a projected area ratio which can be used to estimate mechanical properties of sands.

Mahmood and Mitchell (1974) studied particle orientation as a function of three compaction methods: pouring, dynamic compaction, and static compaction. They found that a nearly random particle orientation was produced by dynamic compaction. The static technique resulted in grain orientations that were nearly 45 degrees from a horizontal plane while the pouring method produced a strong preferred horizontal orientation of particles. Mitchell et al. (1976) studied the effects of sand fabric and sample preparation method on the liquefaction behavior of sand and observed differences in dynamic behavior based on the specimen preparation method used. Microstructural analyses of specimens showed measurable differences in grain orientations and resultant fabrics from the different compaction methods. In addition, Mitchell et al. (1978) investigated the fabric of undisturbed sands from the city of Niigata Japan, the site of a major earthquake in 1964, and also observed the influence of microstructure on dynamic soil behavior..

Using mercury intrusion porosimetry techniques, Juang and Holtz (1986) were able to characterize the effects of compaction and compaction moisture content by a Pore Size Distribution (PSD) index. They observed distinct pore size distributions for the different packing methods investigated. Nimmo and Akstin (1988) found that the permeability of sandy soils was highly dependent upon the compaction method, especially moist compaction. They attributed variations in permeability to preferred grain orientations during compaction as a result of moisture being present as opposed to dry packing which should typically give a random orientation.

These various studies have recognized and described the influence of soil microstructure on soil behavior. However, a clear and concise explanation of the phenomena observed, especially with respect to dynamic behavior, is not currently available and considerable research remains to be done in this area.

IV. Experimental Investigation

The investigation described herein was conducted to examine the stress transmission behavior and development of microstructure in dynamically compacted specimens of unsaturated sand. The SHPB device at RDCM was used to obtain data on dynamic stress transmission characteristics. Compacted specimens were also epoxied and sectioned for microstructural analysis to study the orientation of particles as a function of compaction moisture (saturation). These aspects of the study are outlined in the following sections.

A. Description of Granular Material Tested

A commercially available granular soil designated as "Ottawa 20-30 sand" was used in all tests. Approximately 227 kg (500 lbs) of the sand was obtained from the Ottawa Silica Company and random samples were taken from the bulk quantity for this investigation. The material is a uniformly graded, subrounded to rounded, medium sand with no fines and is classified as SP according to the Unified Soil Classification System (USCS). Physical index properties for the Ottawa 20-30 sand are summarized in Table 1 and a grain size distribution curve is shown in Figure 2.

B. Dynamic Compaction of Specimens

For this study, Ottawa 20-30 sand specimens were dynamically compacted in a thick-walled stainless steel tube to a constant dry density of 1.715 g/cm^3 (107.0 pcf) at varying degrees of saturation (different initial moisture contents). A thick-walled tube was used to simulate the one-dimensional boundary condition encountered in the field near an explosion. The tube was 7.62 cm (6.00 inches) long with an inside diameter of 5.08 cm (2.00 inches) and a wall thickness of 2.54 cm (1.00 inches). The compaction process used a Standard Proctor hammer, ASTM D-698 (ASTM 1990), to consistently apply a controlled amount of compactive effort per impact to each soil specimen (7.5 Joules or 5.5 ft-lbs per impact). Specimens were formed using four individually compacted layers of equal weight such that a final total specimen length of 10.16 cm (4.00 inches) was

obtained. The final compacted thickness of each lift was 2.54 cm (1 inch) and the degree of saturation was varied from near 0% (dry) to about 75% . Figure 3 shows a schematic of a typical specimen being compacted.

TABLE 1. Physical Properties of Ottawa No. 20/30 Sand.

USCS Classification	SP
Specific Gravity	2.65
D ₅₀ particle size	0.70 mm
^a C _u	1.40
^b C _c	1.03
^c Percent passing #100 sieve	<1 %
^d Maximum dry density	1,763 kg/m ³
^d Minimum dry density	1,587 kg/m ³
Maximum void ratio	0.669
Minimum void ratio	0.504
<hr/> Note: ^a Coefficient of Uniformity ^c U.S. Standard sieve size ^b Coefficient of Curvature ^d Data from Ottawa Silica Sand Co.	

At saturations near 0%, each layer of dry soil was poured directly into the tube and then compacted. In preparing moist specimens, the required amount of water for a given degree of saturation (at final compacted density) was added to the originally dry soil, thoroughly mixed in and then allowed to equilibrate before compacting. The amount of compactive effort required to achieve the desired dry density was observed to vary with the amount of moisture present (saturation). Since the specimens ranged in degree of saturation from near 0% to about 75%, the tests were conducted on unsaturated specimens which implies that both continuous air and water phases exist in the soil (eg. pockets of

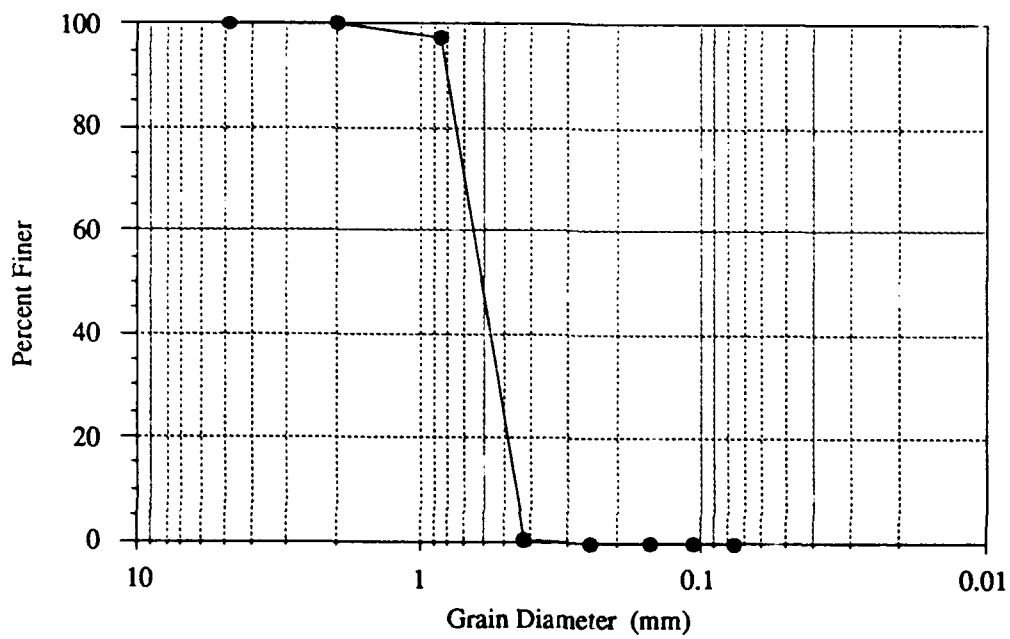


Figure 2. Grain Size Distribution for Ottawa 20-30 Sand.

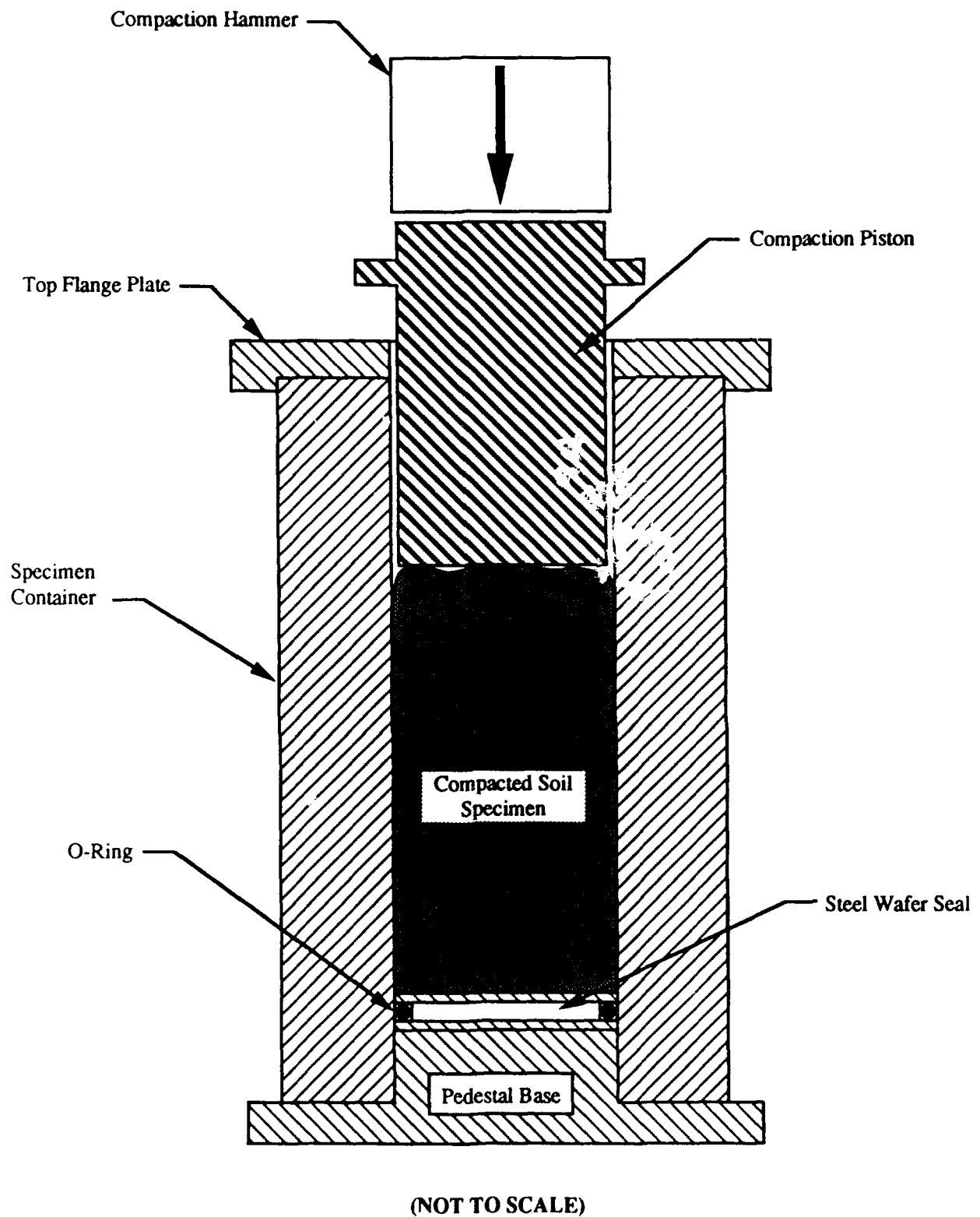


Figure 3. Typical Assembly Drawing for Specimen Container.

occluded air or water are not present). For most soils, this generally occurs at saturations less than about 85% (Corey, 1977).

C. Split-Hopkinson Pressure Bar (SHPB) Tests

The SHPB device has been used by several researchers to study dynamic stress transmission in soils (see for example Charlie et al., 1990; Felice et al., 1987; Ross et. al., 1986; Ross, 1989; Veyera, 1989). The system subjects a specimen to a one-dimensional transient compressive stress wave generated by impacting a steel bar on one side of the specimen with a projectile. An identical steel bar on the other side of the specimen captures the stress wave transmitted by the soil specimen. Transmitted energy and wave speed are determined by analyzing transient measurements from strain gages mounted on the incident and transmitter bars. Details of the SHPB device, principles of operation, and theory are given by Ross (1989). A schematic of a soil specimen in the SHPB device ready for testing is shown in Figure 4.

Two series of dynamic compressive tests were performed on the Ottawa 20-30 sand to study stress transmission characteristics as a function of saturation and microstructure. One series involved compacting specimens moist and then immediately testing them in the SHPB device. The other series involved compacting the specimens moist, carefully drying them in the oven and then testing them in the SHPB. The second series was used to evaluate the importance of the presence of water in the pore spaces during testing and also to determine if the effects produced during moist compaction remained in the microstructure after the moisture is removed.

D. Epoxying and Sectioning of Compacted Specimens

A commercially available bonding agent "EPOTEK 301" was used to preserve the structure of prepared compacted specimens for microstructural analysis. EPOTEK 301 is a high strength, spectrally transparent, low viscosity (100 cps), two component epoxy. Curing time is typically about one hour in a dry, temperature controlled oven at 65 ± 2 °C and specimens can also be cured at room temperature overnight. When

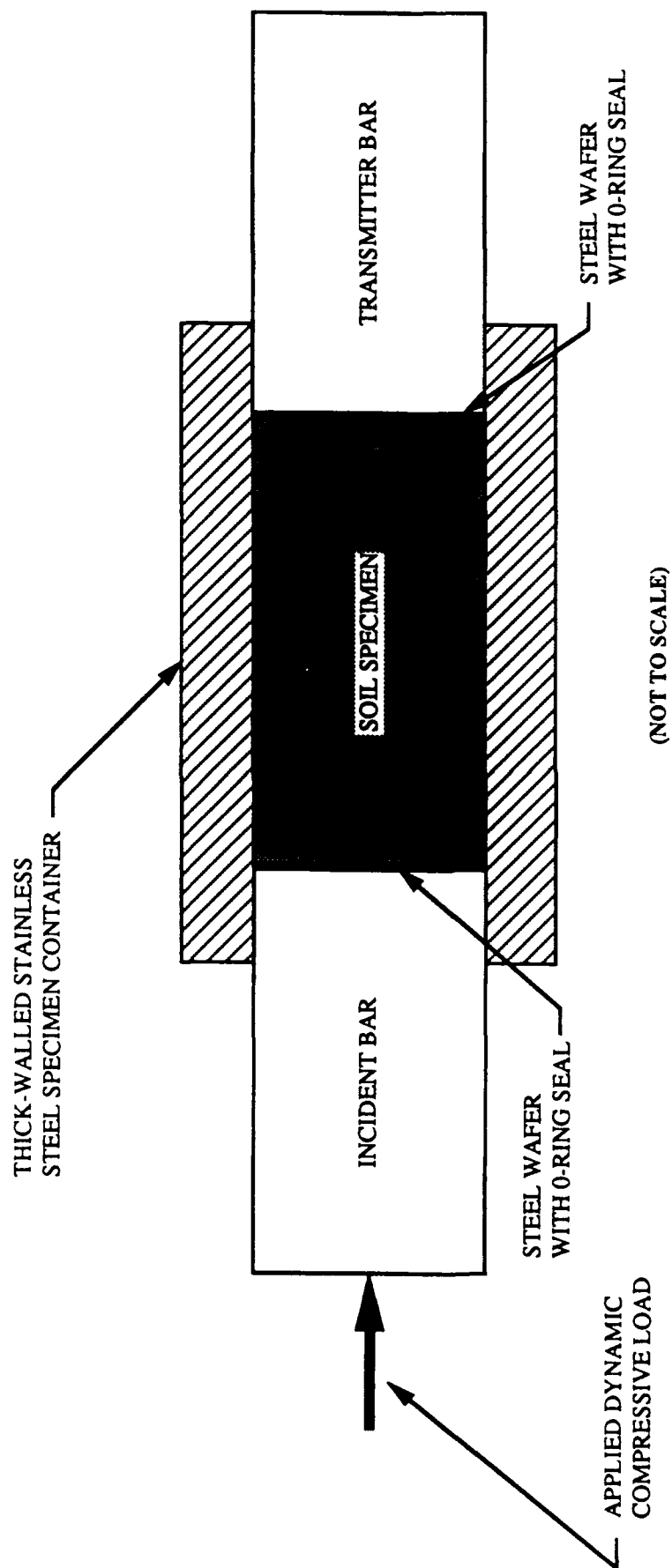


Figure 4. Schematic of Soil Specimen Container Being Tested in the Split-Hopkinson Pressure Bar Device.

fully cured, the epoxy is strongly bonded to the specimen grains and maintains the integrity of the soil structure during cutting and polishing operations. Commercially available biological stains (Sudan IV and Sudan Black) were mixed with the epoxy as a coloring agent to provide adequate contrast between the pores and grains. Details of the laboratory procedure for preparing, epoxying and sectioning compacted sand specimens are given in a previous report (Veyera and Fitzpatrick, 1990), and the general procedure is briefly outlined as follows:

- a) Specimens compacted moist are dried overnight in a temperature controlled oven. Specimens compacted dry are epoxied immediately (step b).
- b) EPOTEK 301 epoxy is carefully introduced into the specimen under atmospheric pressure to saturate the void spaces. The specimen is then cured in a temperature controlled oven at 65 ± 2 °C for about 1 hour.
- c) After curing, the specimen is extruded from the stainless steel container.
- d) The sectioning process involves 2 major steps with several sub-steps:
 - i) Initial Rough Cutting - 1) Preliminary cutting of specimen; and 2) Preparation of vertical and horizontal sections for grinding and polishing;
 - ii) Grinding and Polishing - 1) Initial coarse grinding; 2) Rough polishing; and 3) Fine finish polishing.

The procedure outlined was used to obtain representative sections along both the specimen vertical longitudinal axis and perpendicular to the longitudinal axis at specimen mid-height. Locations of the individual cut sections are shown in Figure 5. After final polishing, the specimen sections are ready for two-dimensional microstructural analysis.

E. Microstructural Analysis

1. Two-Dimensional Imaging

After the epoxied specimens have been sectioned, a two-dimensional image of the grain and void arrangements was obtained for microstructural analysis using photomicrographic techniques. A 35 mm camera with special attachments was mounted on

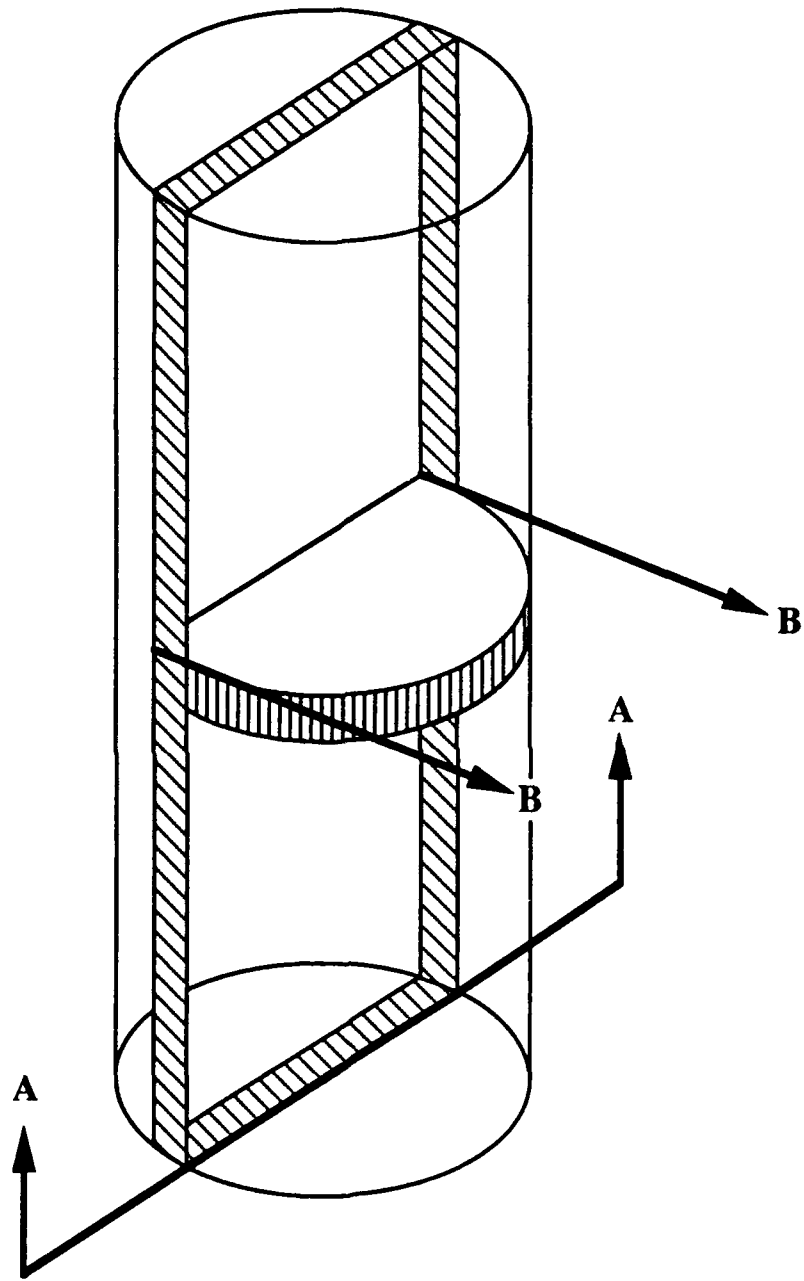


Figure 5. Location of Cutting Planes for Epoxied Specimens.

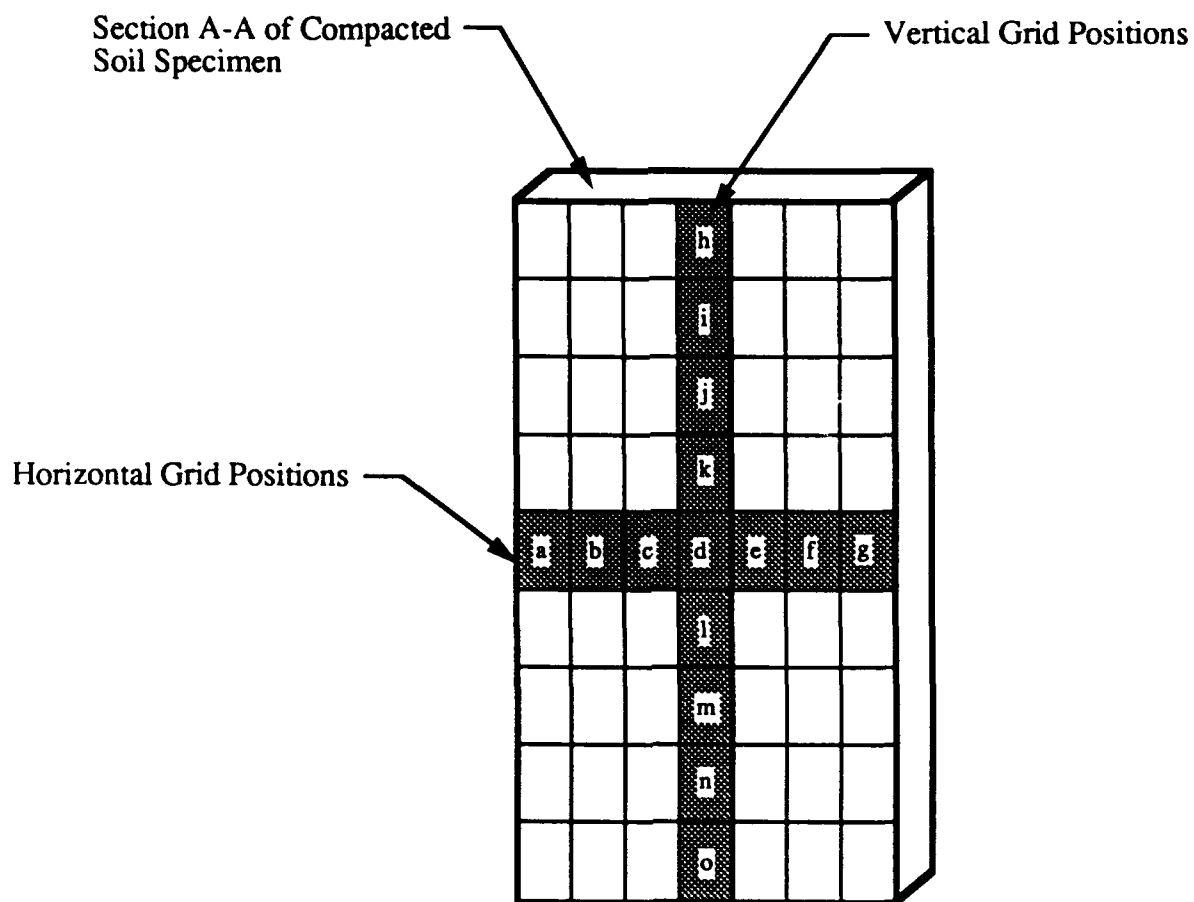
a reflected light microscope with direct lighting used to illuminate specimen cut sections from above. Each specimen section was carefully leveled and centered on the microscope platen. A magnification of 10x was used on all specimen sections. Standard 35 mm color slide film (100 ASA) and a shutter speed 1/60 second were used when photographing the cut section surfaces, which provided optimum visual contrast between grains and epoxy.

Each cut section was systematically photograph according to a prescribed grid position system (see Figure 6). A precise record of frame number and corresponding grid position was maintained for proper identification of slides during analysis. For the purposes of this study, only five grids positions on the vertical cut section (k, d, l, c, and e) were photographed for specimens compacted at each saturation of 0, 18.7, 37.4, 56.2 , and 75% (The remaining grid positions are currently being analyzed.). Two-dimensional particle measurements were made by projecting the slides onto a sheet of paper and tracing individual particles. This data was then used to analyze the microstructure of the compacted specimens. Axial ratio and particle long axis orientation characterization parameters were determined following procedures presented by Oda (1972a) and Campbell (1985), respectively.

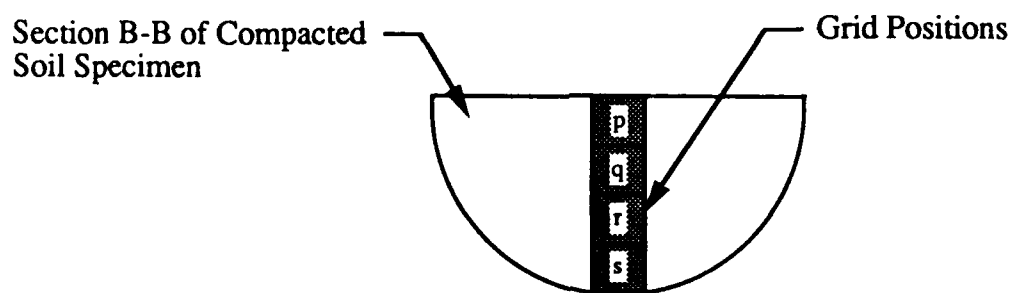
2. Determination of Axial Ratio

A shape factor for studying individual grain particles is the axial ratio, n , which is obtained by taking measurements from a two-dimensional section of a particle. The axial ratio is an indication of how closely a particle is to being spherical in shape and can range from 0 to 1. For example, a ratio of 0.9 would indicate a particle shape approaching that of a sphere (1.0 would be a sphere), whereas a ratio of 0.1 would indicate an elliptical or elongated shape. The procedure suggested by Oda (1972a) was followed:

- 1) determine the length of the particle's short axis, L_2 , by finding the diameter of the largest possible circle that can be inscribed on the particle (see Figure 7a),

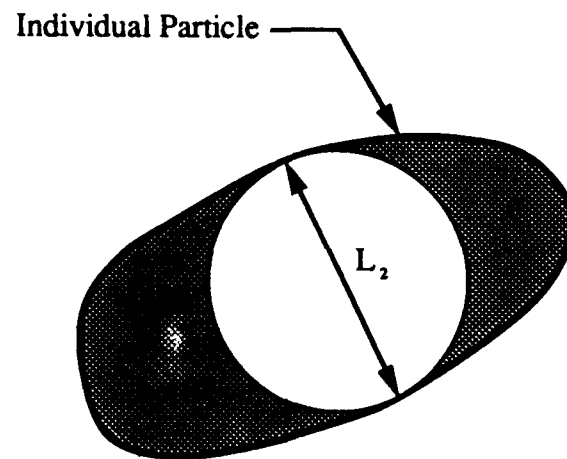


a. Grid Positions for Vertical Specimen Cut Sections.

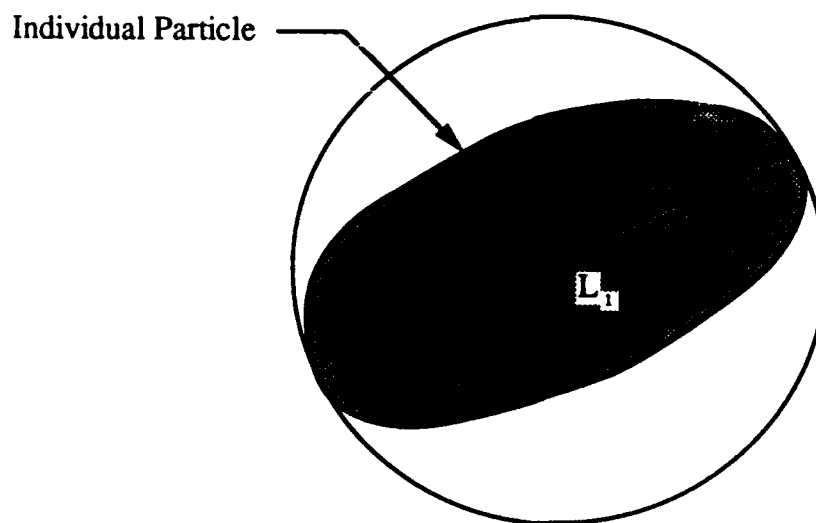


b. Grid Positions for Horizontal Specimen Cut Sections.

Figure 6. Location of Specimen Grid Positions for Photographic and Microstructural Analysis.



a. Short Axis (L_2) Determination



b. Long Axis (L_1) Determination.

Figure 7. Determination of Axial Ratio Parameters for an Individual Particle.

2) determine the length of the particle's long axis dimension, L_1 , by finding the diameter of the smallest possible circle that can be circumscribed on the particle (see Figure 7b),

3) the axial ratio for an individual particle, n , is then:

$$n = \frac{L_2}{L_1} \quad (\text{Eq. 1})$$

4) the axial ratio for a group of particles, n_{avg} , is then:

$$n_{\text{avg}} = \left(\frac{1}{f}\right) \sum_{i=1}^{i=f} \left(\frac{L_2}{L_1}\right)_f \quad (\text{Eq. 2})$$

Axial ratio data are useful in classifying the general shape of particles and are also an indication of the potential arrangement of particles that can be expected from compaction (ie., the closer to being spherical in shape, the more random the arrangement). In addition, some physical properties of sands have been correlated with axial ratio results (Oda, 1972a). Table 2 provides rankings and visual descriptions of particle shapes based on axial ratio data (roundness ratio), which would classify Ottawa 20-30 sand as consisting of well rounded particles.

3. Determination of Particle Long Axis Orientation

Particle long-axis orientation is a standard method used in the microstructural analysis of granular soils. The long-axis orientation of individual particles was evaluated based on the method developed by Campbell (1985). The technique involves determining both longest axis, L_1 , and its shortest axis, L_2 , of a particle. A rectangle is then constructed around the grain and the orientation angle, θ , made by the

long-axis direction with respect to a set of Cartesian coordinate system reference axes is measured (see Figure 8). For each grid position, orientation measurements (class intervals) were made at 10° intervals and used to construct frequency histograms of the data. The results then provide information on any preferred orientation of particles that may occur locally in a specimen or globally for the specimen as a whole. Correlations between compaction method (pluvial, static, or dynamic) and soil microstructure can also be related to preferred particle orientation data.

TABLE 2. Description of Roundness Classes (Pettijohn, 1957).

Class Name	Roundness Value ^a	Description
Angular	0 - 0.15	Strongly developed faces with sharp edges and corners; secondary corners ^b are numerous.
Subangular	0.15 - 0.25	Strongly developed faces with somewhat rounded edges and corners; secondary corners are numerous.
Subrounded	0.25 - 0.40	The edges and corners are rounded and the area of flat faces is comparatively small; secondary corners are much rounded and reduced in number (5 - 10).
Rounded	0.40 - 0.60	Flat faces are practically absent; all edges and corners are rather broad curves, and there may be broad re-entrant angles; secondary corners have disappeared.
Well-Rounded	0.60 - 1.00	There are no flat faces; the entire surface consists of broad curves.

Note: ^aRoundness value is equivalent to axial ratio.

^bSecondary corners are the many minor convexities seen in the grain profile.
Primary corners are the principal interfacial edges and are fewer in number (3-5).

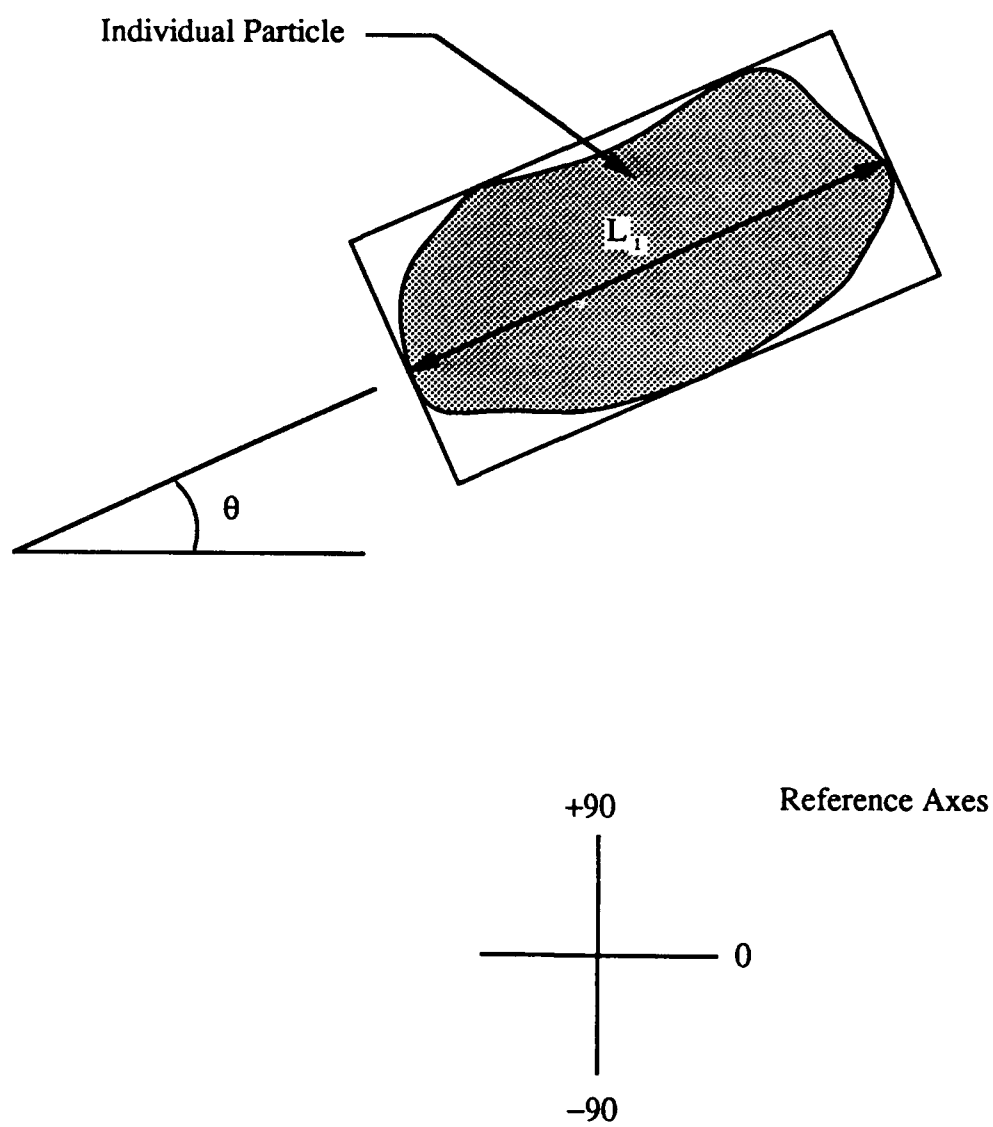


Figure 8. Determination of Particle Long Axis (L_1) Orientation Angle θ .

V. RESULTS and DISCUSSION

A. Dynamic Compactive Energy

A record was kept during compaction of the number of blows required to achieve a dry density of 1.715 g/cm³ (107.0 pcf) at each saturation for each specimen. Figure 9 shows the average compactive energy as a function of saturation (based on results from 47 compacted specimens). The data has been normalized to the average total compactive energy required at 0% saturation for the first layer. The results show that the required compactive energy increases from 0% to about 20% saturation, remains constant from about 20% to 50% saturation, and then decreases thereafter.

Figure 10 shows the compactive energy data normalized to the average compactive energy required at 0% saturation for each individual layer. The greatest compactive effort was always applied to the first layer regardless of moisture present. The data suggest that the required compactive energy does not decrease at higher saturations. It may be that at higher saturations the pore water is more effective in resisting the compaction energy in a single layer, particularly since water is unable to migrate to other layers. There also may be a partial loosening of the soil in this first layer due to the stress wave reflected from the pedestal during compaction. The second and third layers require about the same amount of compactive effort and show a reduction in required compaction energy above about 50% saturation. The fourth and final layer required the least amount of compactive effort but still exhibited a dependency on moisture. Also, layers 2, 3 and 4 show a trend with increasing saturation similar to that shown in Figure 9.

From the results shown in Figures 9 and 10, it can be seen that there is a strong dependency of compactive energy on moisture for a constant dry density packing. This can be attributed to variations in overall specimen stiffness with moisture which may be due to the formation of preferred particle orientations and the presence of capillary pressures during compaction.

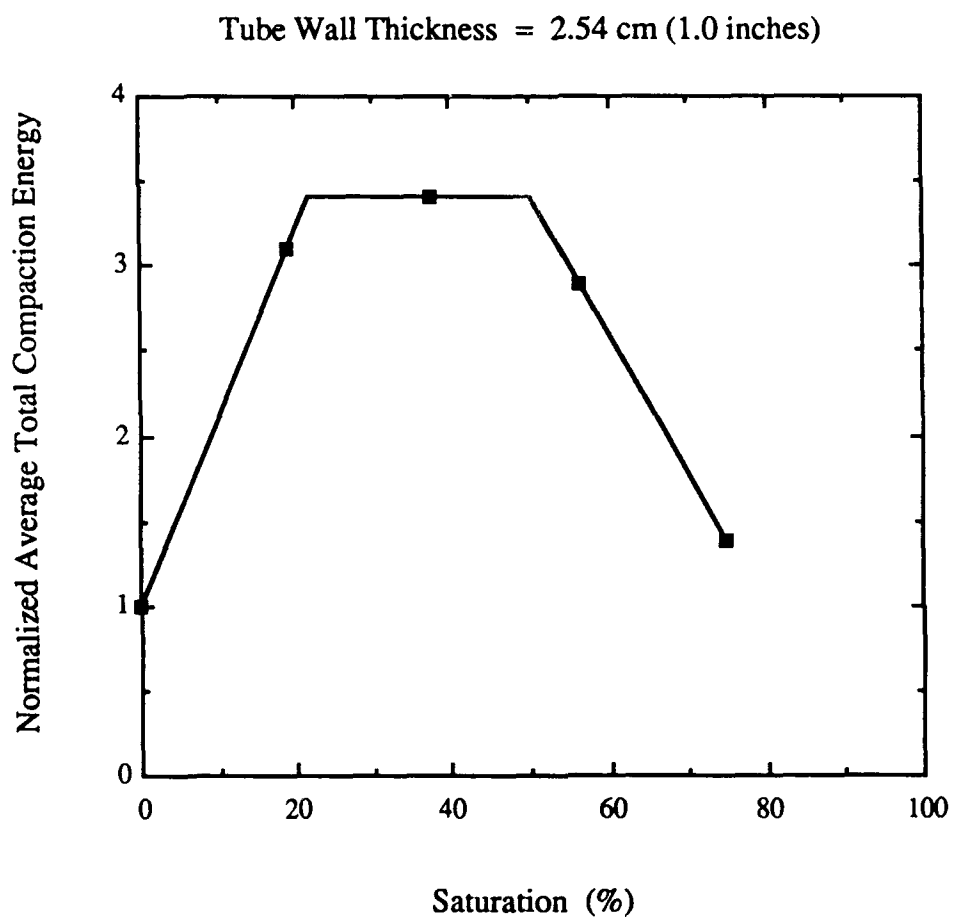


Figure 9. Normalized Average Total Compactive Energy for Ottawa 20-30 Sand Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

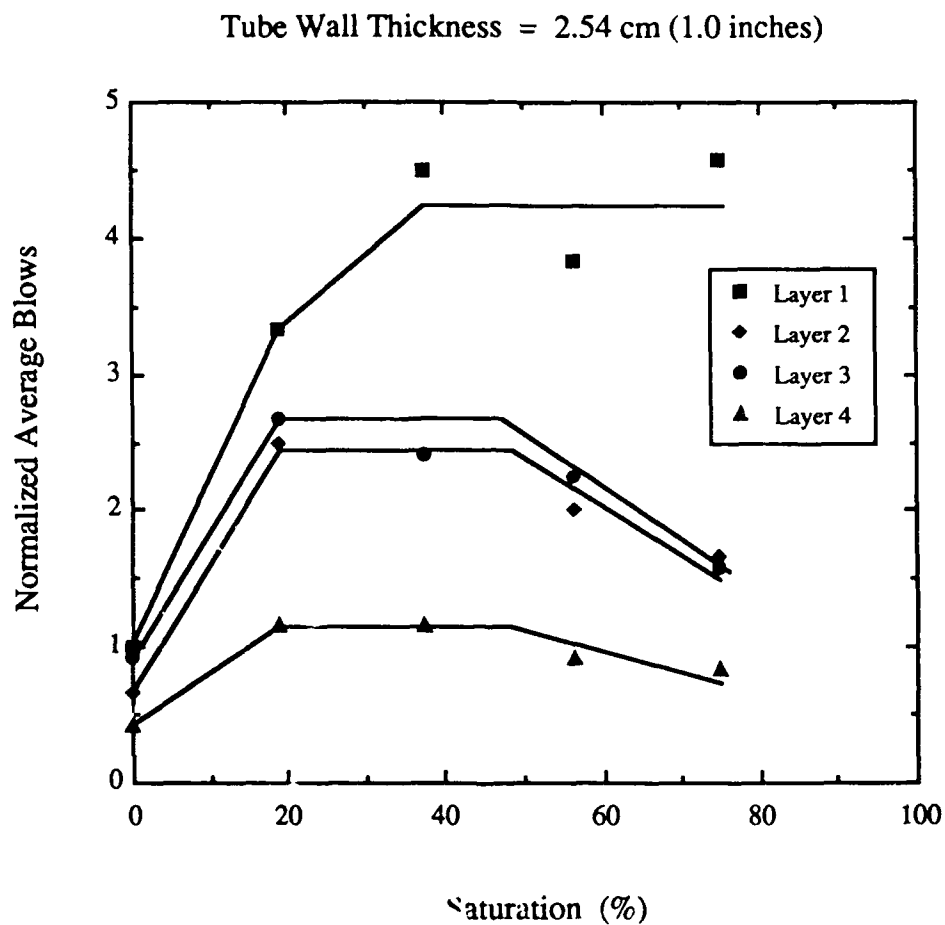


Figure 10. Normalized Average Compactive Energy as a Function of Saturation by Layer for Ottawa 20-30 Sand Compacted to Final Dry Density of 1.715 g/cc (107.0 pcf).

B. Split-Hopkinson Pressure Bar (SHPB) Tests

Two series of dynamic compressive tests using the SHPB were performed on Ottawa 20-30 sand to study stress transmission characteristics as a function of saturation and microstructure. One series involved compacting specimens moist and then testing them immediately to provide information about the effect of moisture during compaction and testing on the dynamic soil response. The second series was conducted on specimens compacted moist, oven-dried, and then tested in the SHPB. These tests were used to determine if the conditions developed during compaction remained locked in the soil structure even after the moisture has been removed from the pores. Care was used in handling these specimens so that the structure formed during compaction would not be disturbed prior to testing.

The data shown in Figures 11 and 12 represent average normalized values obtained from 24 specimens compacted moist and tested moist, and 23 specimens compacted moist and tested dry from this study. In addition, the results from 35 specimens compacted moist and tested moist by Ross (1989) are also shown. The data in the figures have been normalized to the average value at 0% saturation from each study. Ottawa 20-30 sand specimens and identical loading and boundary conditions were used in each investigation. The only difference was in the compacted dry densities which were 1.715 g/cc (107.0) in this study, and 1.750 g/cc (109.2 pcf) in Ross's investigation.

Figure 11 shows the variation in measured dilatational wave speed normalized to the value at 0% saturation. The wave speed is a measure of the compressive wave propagation velocity through the compacted soil and is a function of material density and stiffness. For specimens compacted moist and tested moist, the wave speed increases from 0% to about 20% saturation, remains constant from about 20% to 50% saturation, and then decreases thereafter. The data from Ross (1989) show a similar trend; however, the wave speed magnitudes are lower which was not expected since the dry density (stiffness) was higher. The differences may be due to the fact that velocities are based on arrival times

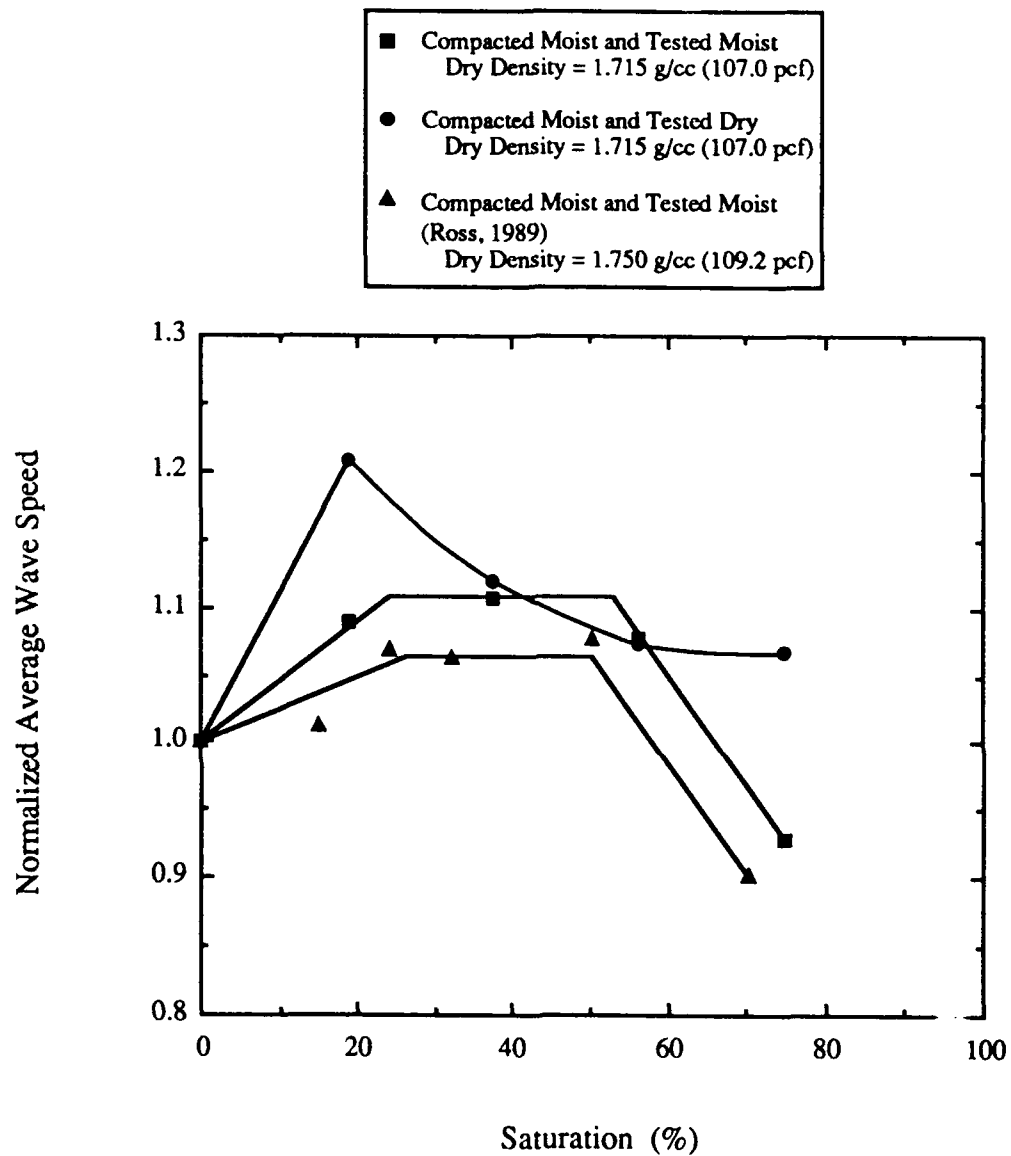


Figure 11. Normalized Average Wave Speed as a Function of Saturation for Ottawa 20-30 Sand.

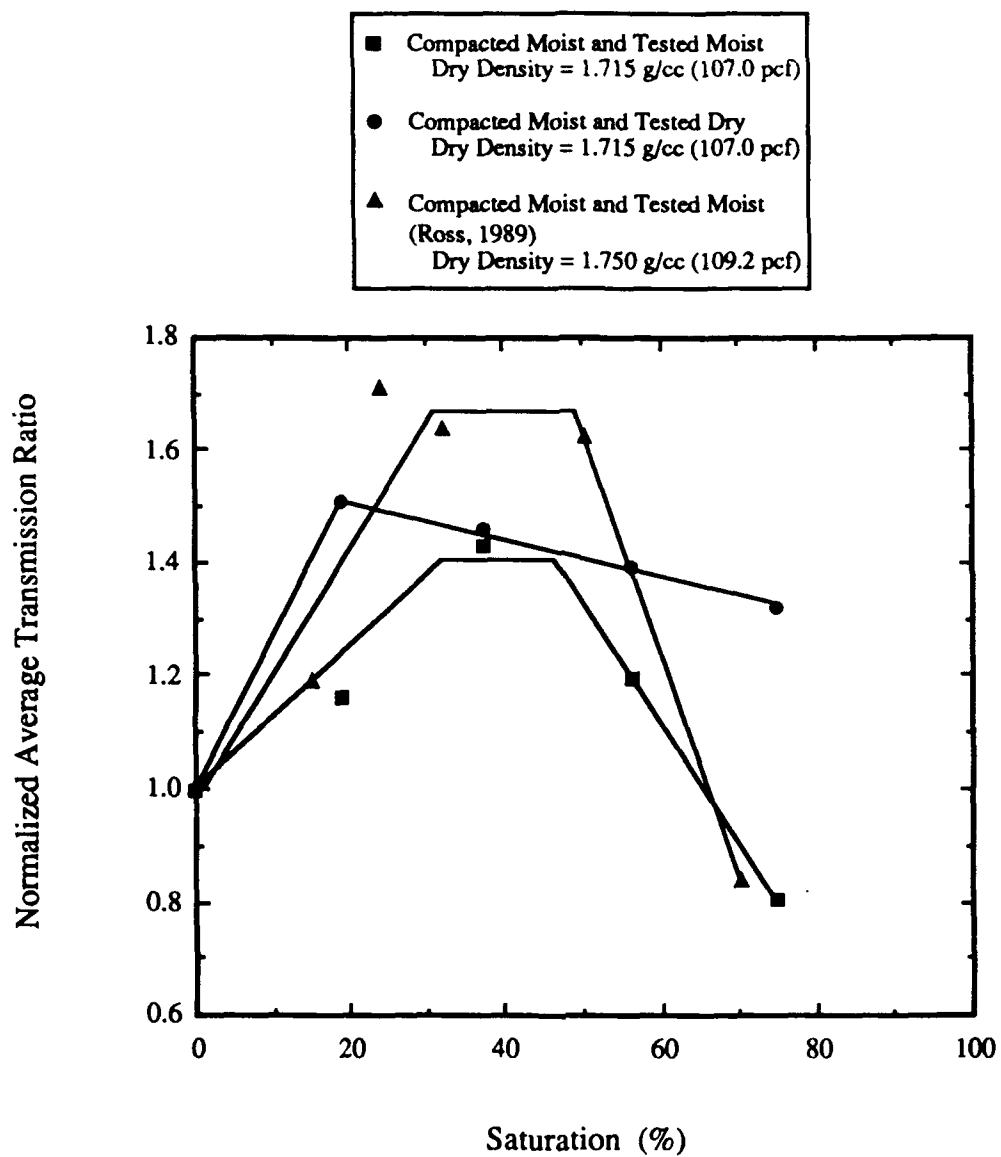


Figure 12. Normalized Average Transmission Ratio as a Function of Saturation for Ottawa 20-30 Sand.

which must be interpreted from the recorded strain gage time-history data. In any event, the same behavior is evident.

An interesting feature in the normalized wave speed data is that magnitudes above about 60% saturation are lower than those at 0%. This implies that the presence of moisture during testing may have a lubricating effect as the compressive energy passes through the specimen. Particles would then be able to move somewhat more freely relative to each other and absorb more energy.

Figure 11 also shows the normalized wave speed data for specimens compacted moist and tested dry. The saturations associated with this data represent the moisture conditions during compaction. The data suggests a trend similar to that previously described; however the wave speeds are generally larger particularly at about 20% saturation and decrease with increasing saturation. This indicates larger specimen stiffnesses compared with those tested moist at the same saturations. In addition, no plateau region exists and the wave speed remains above that at 0% saturation regardless of the moisture condition during compaction. Therefore, it appears that soil microstructural characteristics developed during compaction that influence stress wave propagation velocity remain intact even after the moisture in the pores has been removed.

Figure 12 shows the variation in stress transmission ratio normalized to the value at 0% saturation. The transmission ratio is a measure of the impact energy transmitted by the soil specimen after loading. For specimens compacted moist and tested moist, the transmission ratio increases from 0% to about 30% saturation, remains constant from about 30% to 50% saturation, and then decreases thereafter. The data from Ross (1989) show a similar trend; however, the transmission ratio magnitudes are higher which was expected due to the greater dry density (stiffness). Since these results are based on peak values from the time-history records, no interpretation of the data was necessary and the results therefore, appear to be more consistent with anticipated trends. The results in Figure 12

further show that the transmitted energy at higher saturations (above about 65%) is less than at 0% saturation which may be due to a lubricating effect as previously noted.

Figure 12 also shows the normalized transmission ratio data for specimens compacted moist and tested dry. The data suggests a trend similar to that previously described; however the transmission ratios are larger and only gradually decrease with increasing saturation. As with the wave speed data, this also indicates larger specimen stiffnesses compared with those tested moist at the same saturations. There is no plateau region and the transmission ratio remains above that at 0% saturation regardless of the moisture condition during packing. These results also support the idea that the soil microstructural characteristics developed during compaction which influence stress transmission, remain intact even after drying.

C. Microstructural Analysis

Figure 13 shows the results of the axial ratio particle analysis and is a composite plot of over 2500 individual particles which were manually analyzed. The results indicate that for these particles the average axial ratio is 0.674 (standard deviation of 0.106) which means that the particles are approaching spheres in shape. Based on the axial ratio data obtained, the Ottawa 20-30 sand would be classified as consisting of well rounded particles according to the criteria shown in Table 2.

Figures 14 through 28 show the results of the particle orientation analyses for 5 different grid positions k, d, l, c and e (see Figure 6) at saturations of 0, 18.7, 37.4, 56.2 and 75%. These grid positions are located in the vertical plane (section A-A, Figure 5) at the center of the specimen away from the boundaries. Each grid position contains approximately 100 individual Ottawa 20-30 sand particles. The entire grid is currently being analyzed for both the horizontal and vertical cut sections and the results will be forthcoming in a future report.

The composite results for grid positions c, d, e which extend horizontally across the cut section, are shown in Figures 14-18, and those for grid positions k, d, l which extend

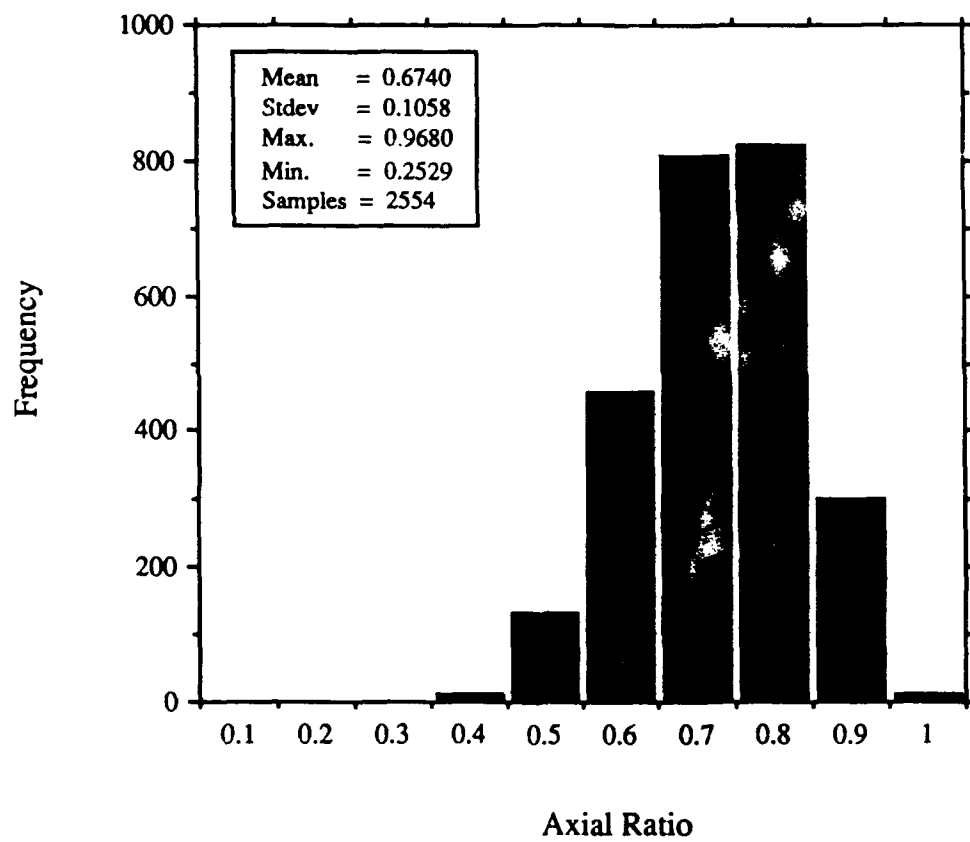


Figure 13. Axial Ratio Histogram for Ottawa 20-30 Sand Particles.

vertically across the cut section are shown in Figures 19-23. The data shown in these figures do not suggest any particular preferred orientation along the directions examined for either grid position grouping.

Considering the extent of the data analyzed to date, it is probably more useful to view the results as a whole and look at a composite representation of all grid positions at each saturation. These results are shown in Figures 24-28 and Table 3 is a summary of the maximum frequency of particle orientations based on the three peak values obtained from each figure. The general trend of the data shows that, for the most part, some preferred grain orientations occur between about -40° and $+20^{\circ}$. However, there does not appear to be a strongly preferred orientation at any saturation. This may be due to the relatively small number of particles examined so far, and these results may change when the remaining data are analyzed. Also, the axial ratio determined for Ottawa 20-30 sand was 0.674 which means that the particles are approaching a spherical shape, and therefore, may not exhibit any significant degree of preferred orientation. This would be consistent with the findings of Oda (1972a) who observed that for axial ratios at or above 0.70, particles tend to pack at random orientations. If this turns out to be the case for Ottawa 20-30 sand once the remaining data has been analyzed, then the influence of other mechanisms, such as locked in stresses and stress anisotropy will need to be considered. At this point, it is difficult to develop any correlations between the microstructural analysis data and the results of the SHPB tests.

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l

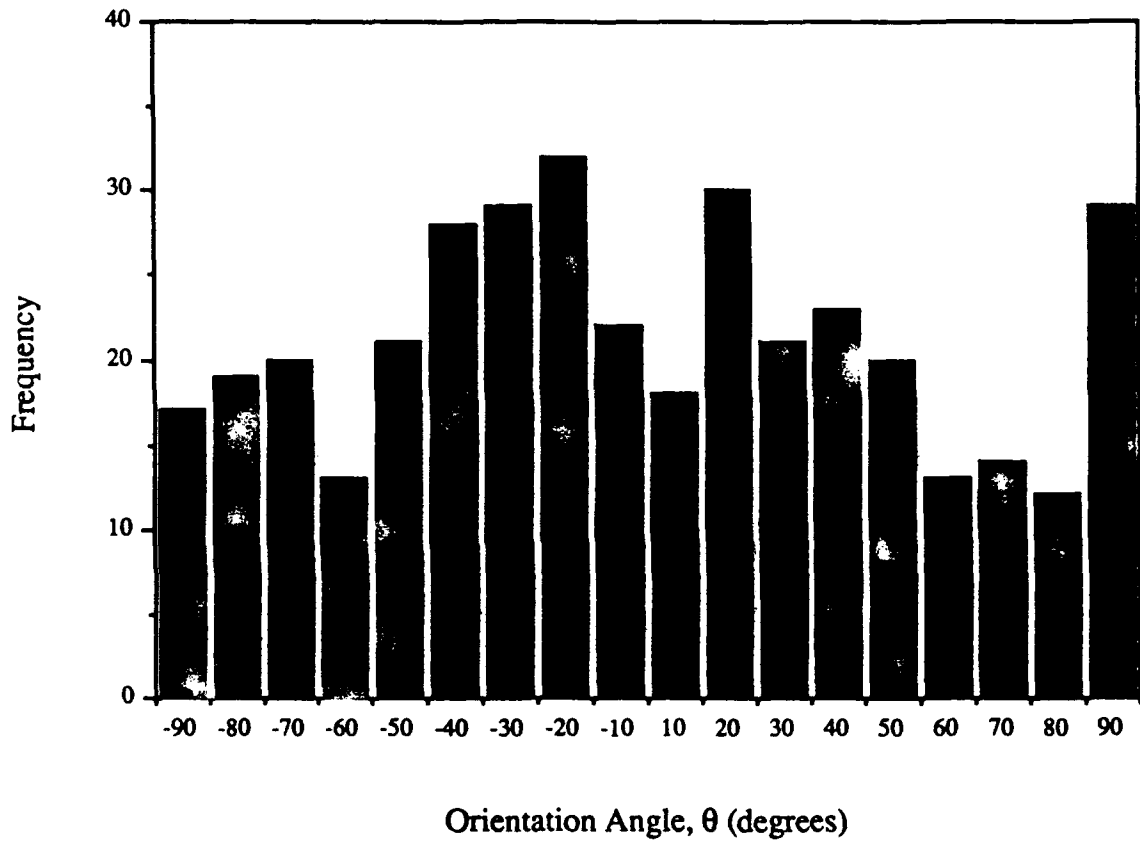


Figure 14. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 0\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l

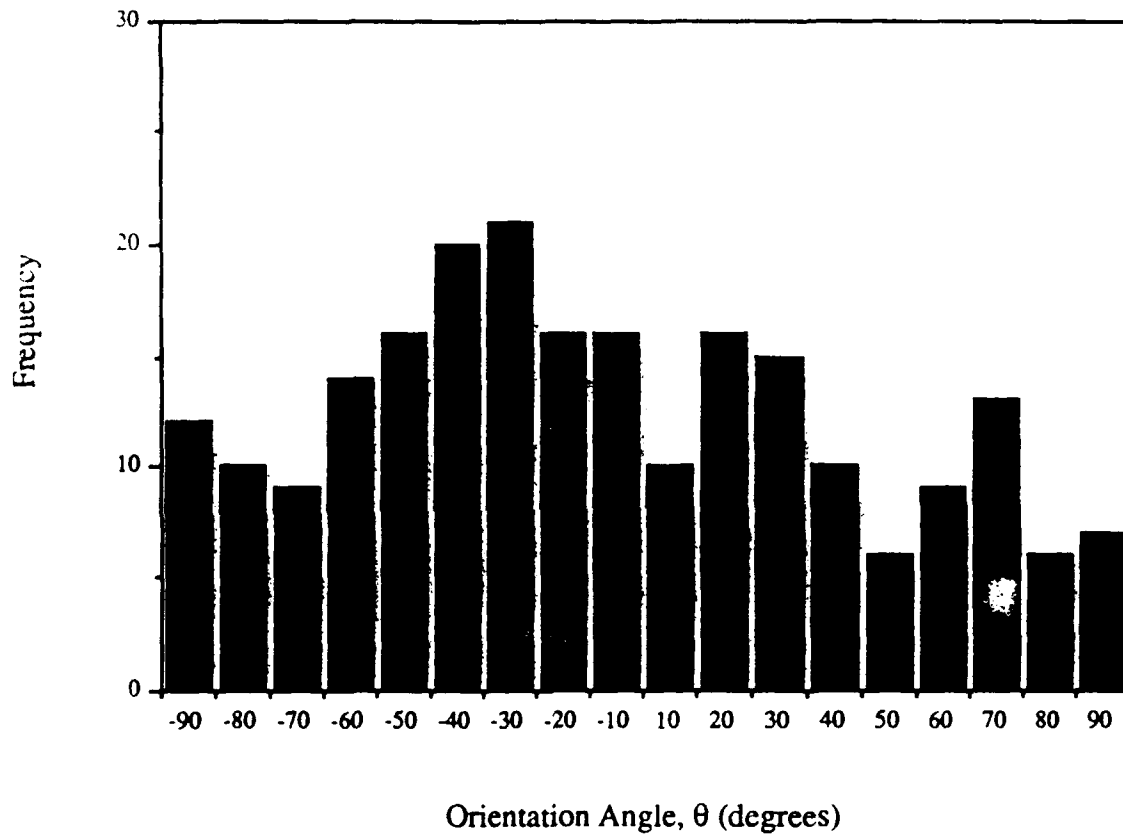


Figure 15. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 18.7\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l

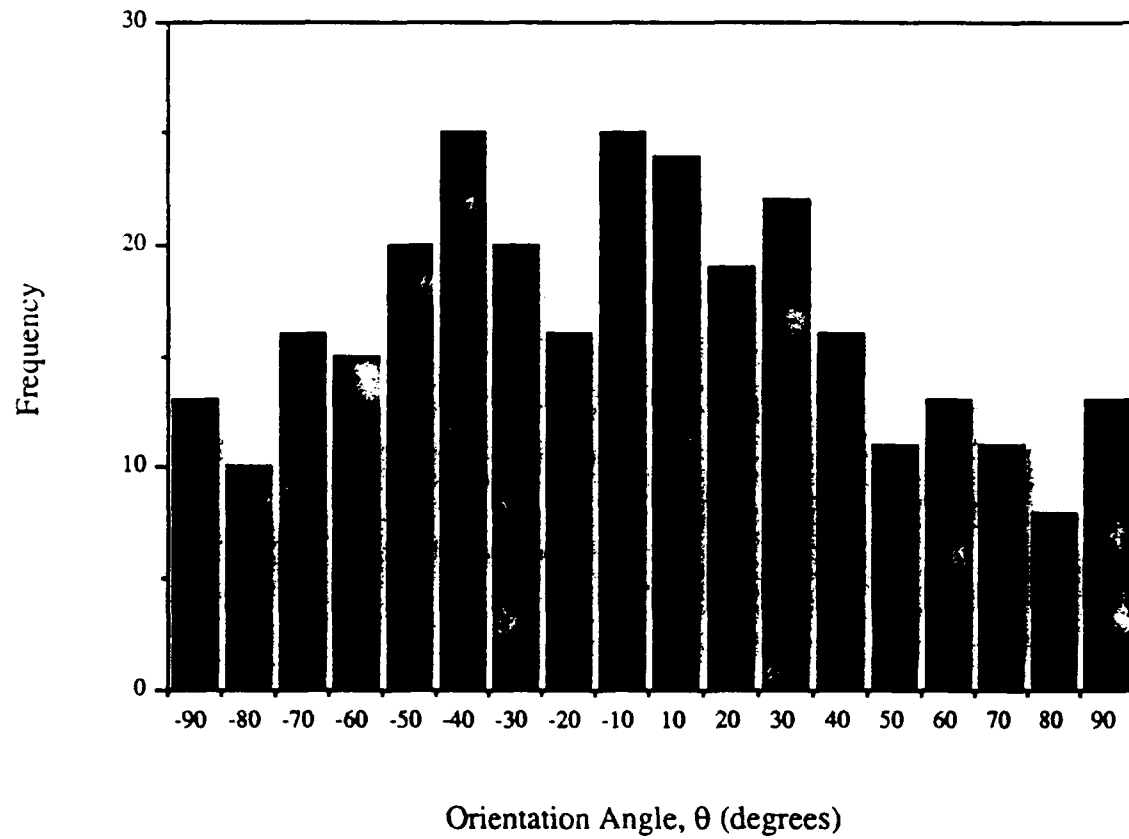


Figure 16. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 37.4\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l

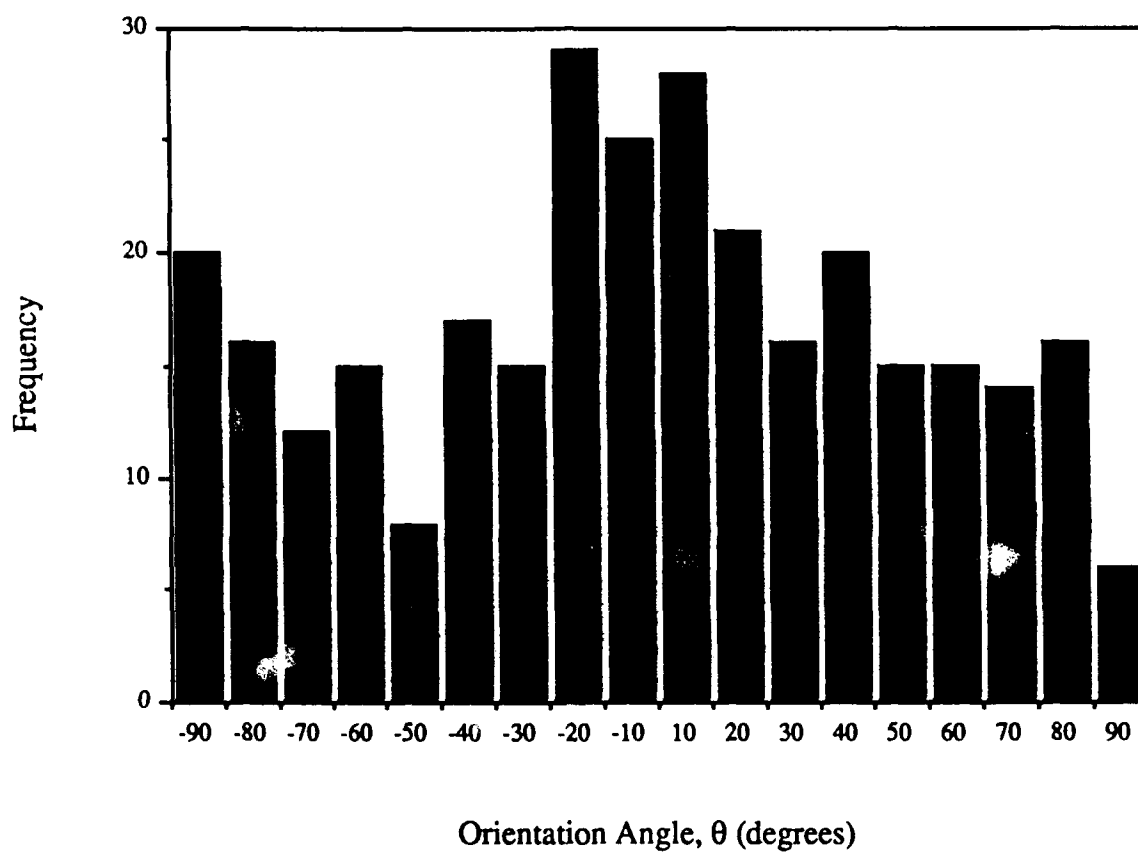


Figure 17. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 56.2\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l

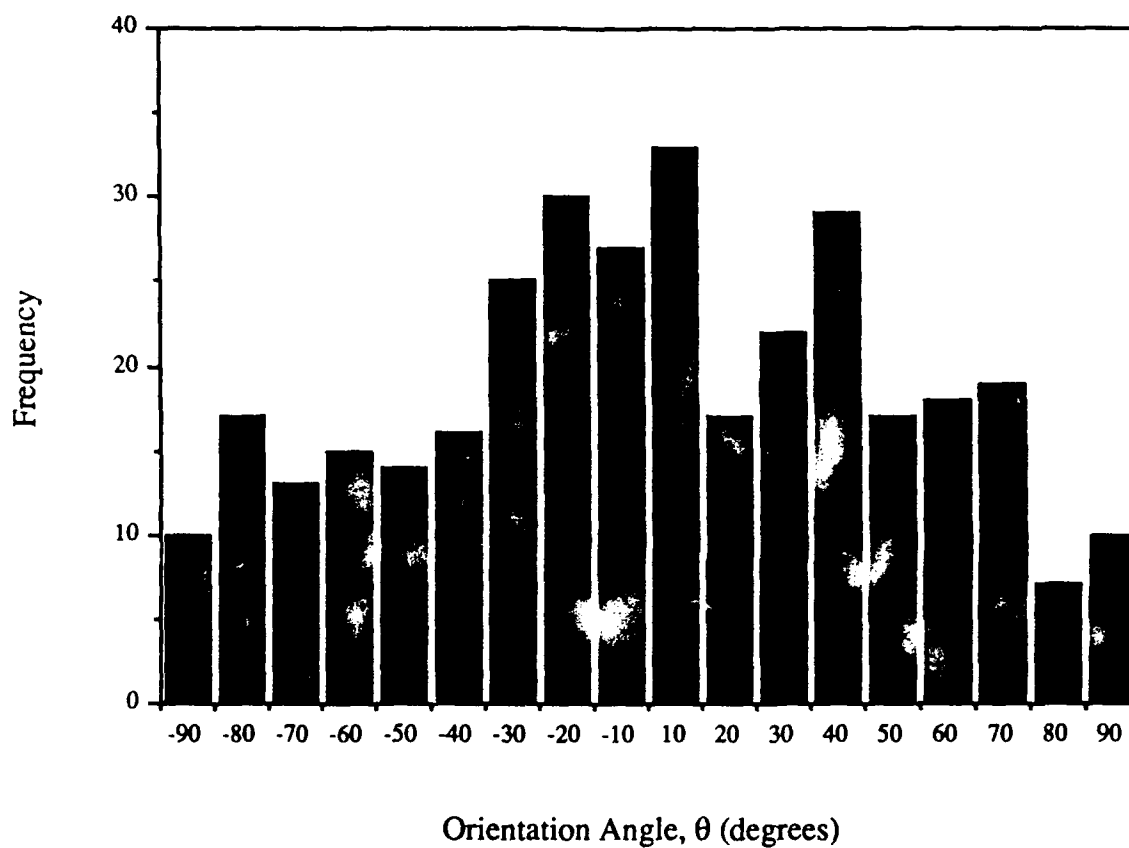


Figure 18. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 75.0\%$ Compacted to a dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = c, d, e

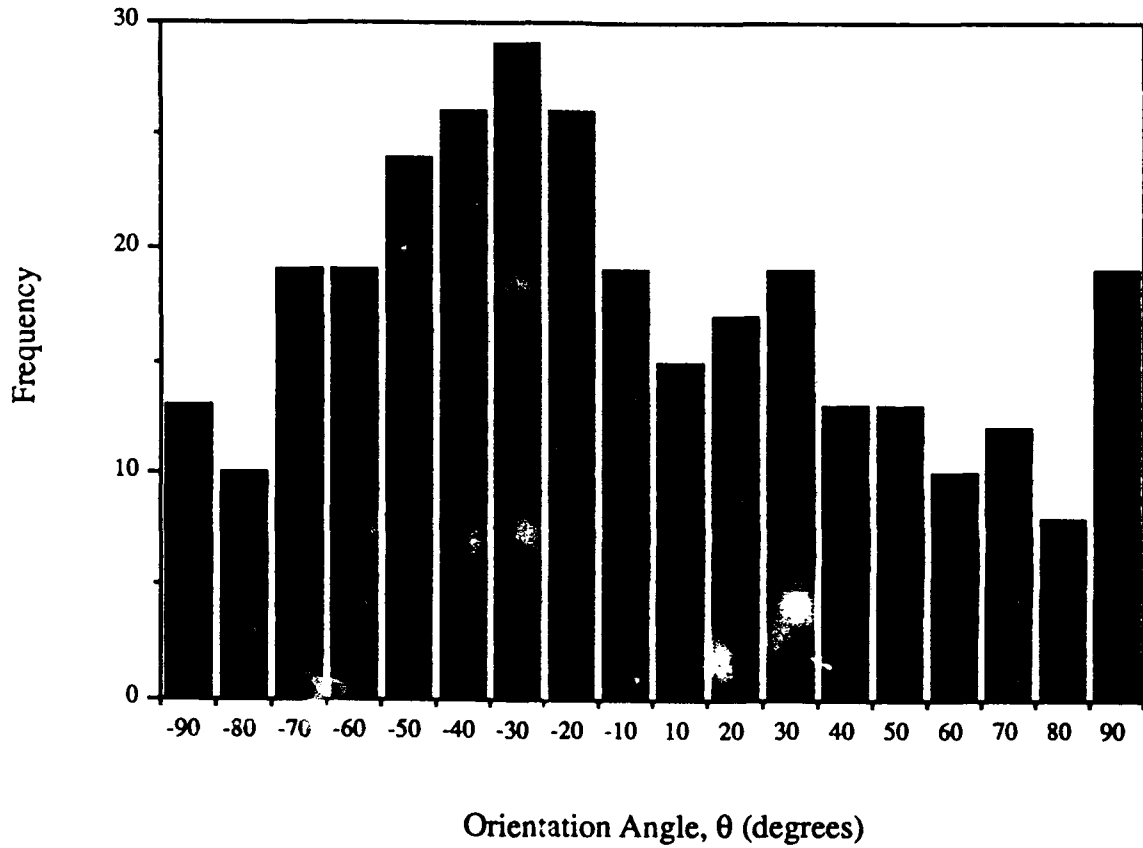


Figure 19. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 0\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = c, d, e

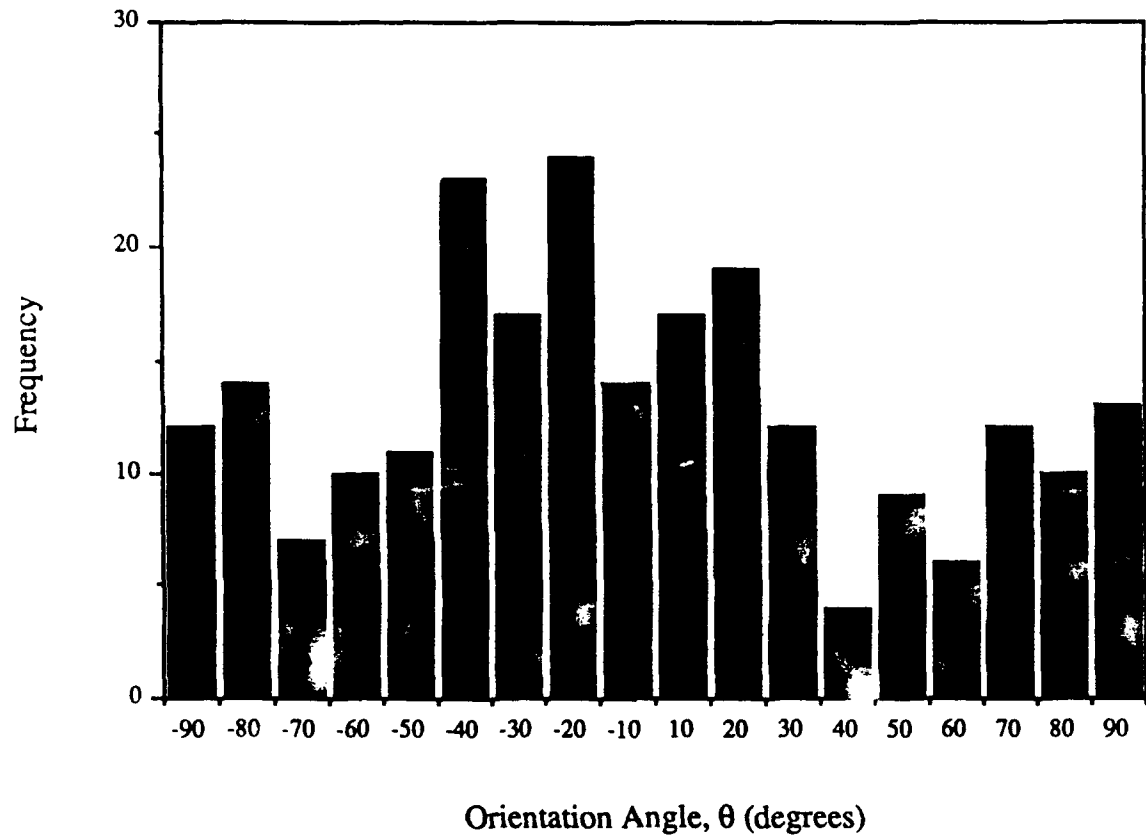


Figure 20. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 18.7\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = c, d, e

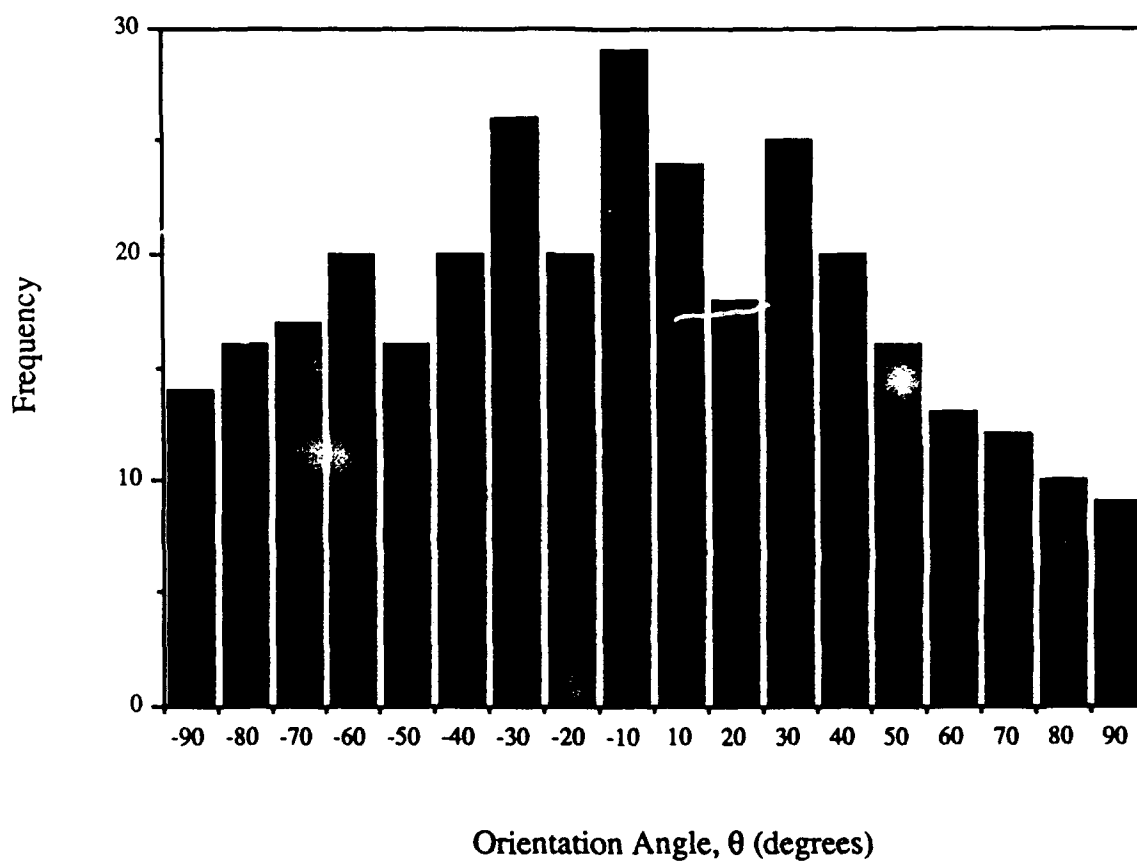


Figure 21. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 37.4\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = c, d, e

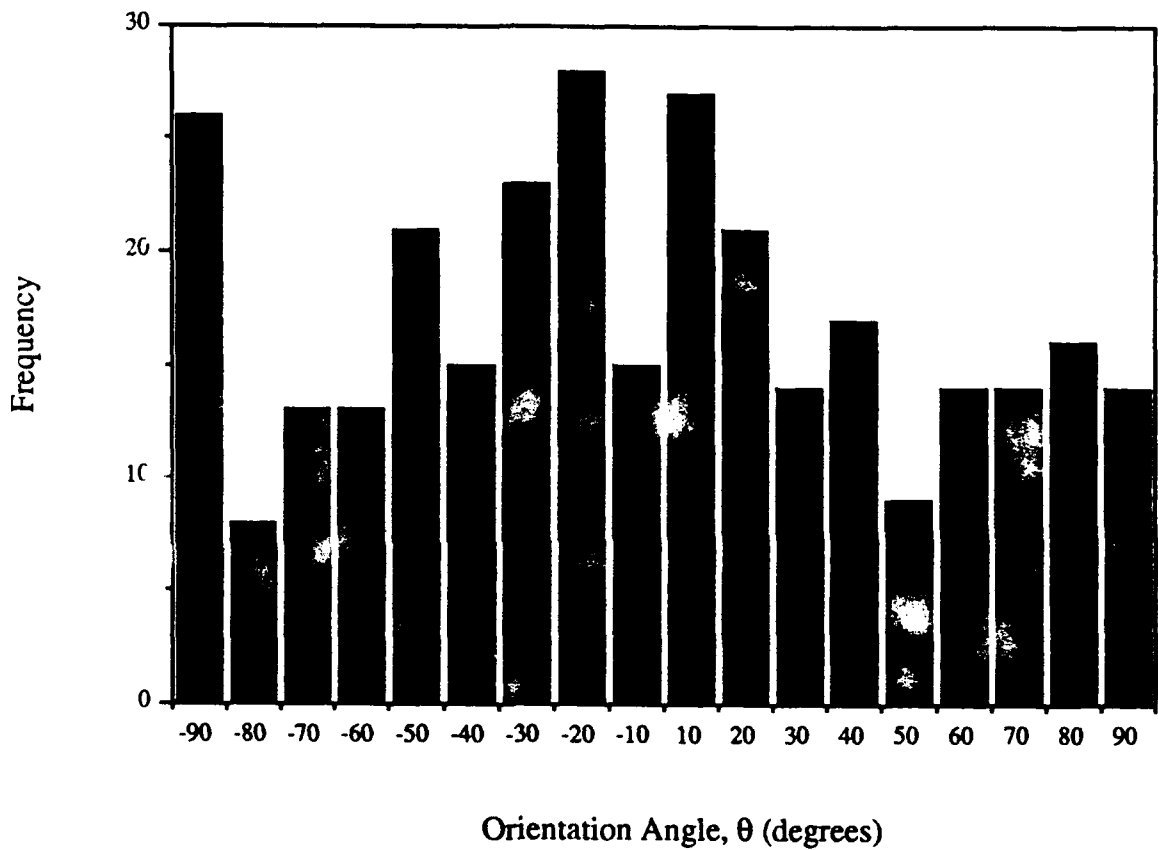


Figure 22. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 56.2\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = c, d, e

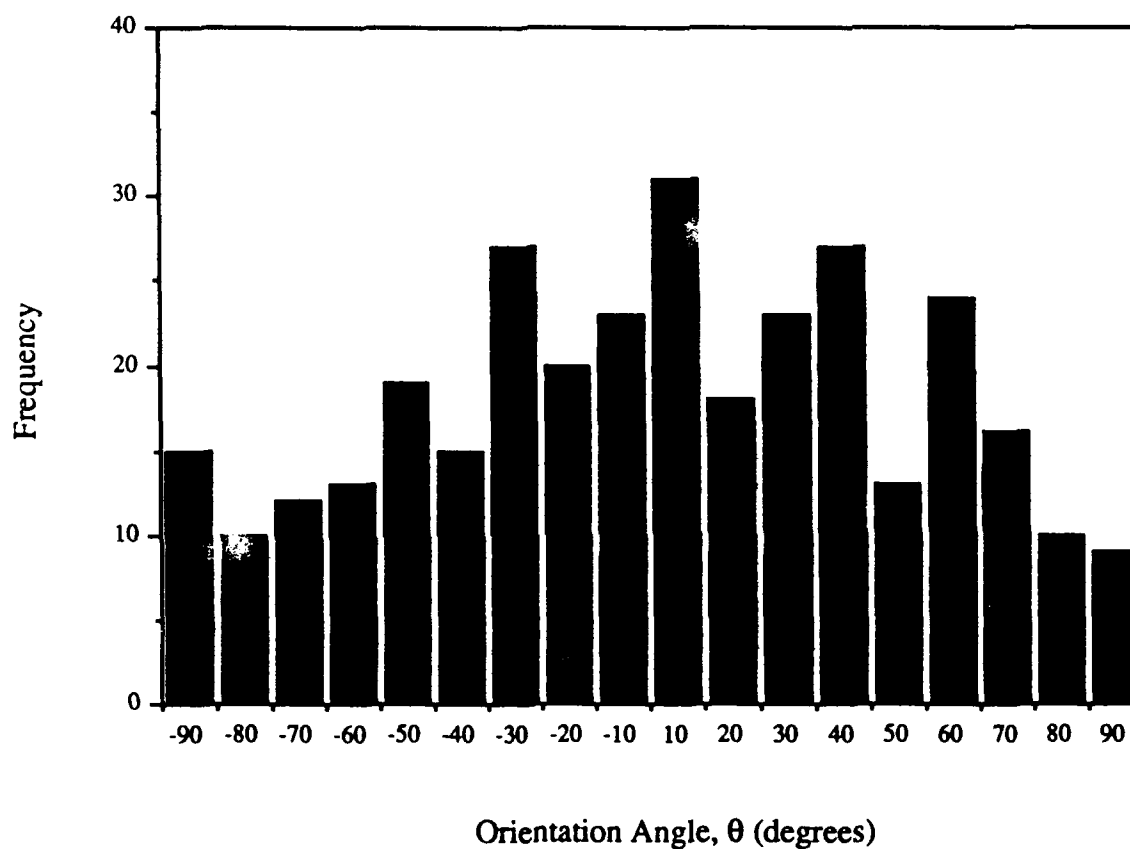


Figure 23. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 75.0\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l, c, e

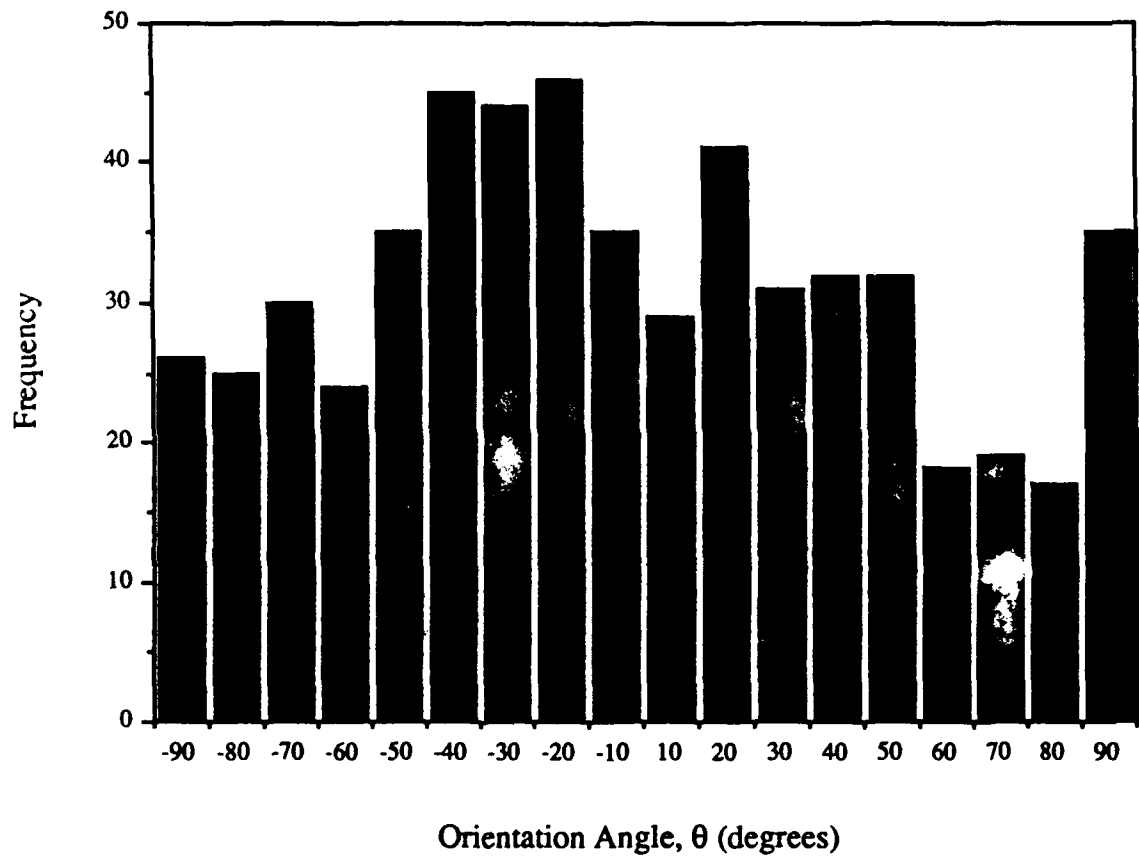


Figure 24. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 0\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l, c, e

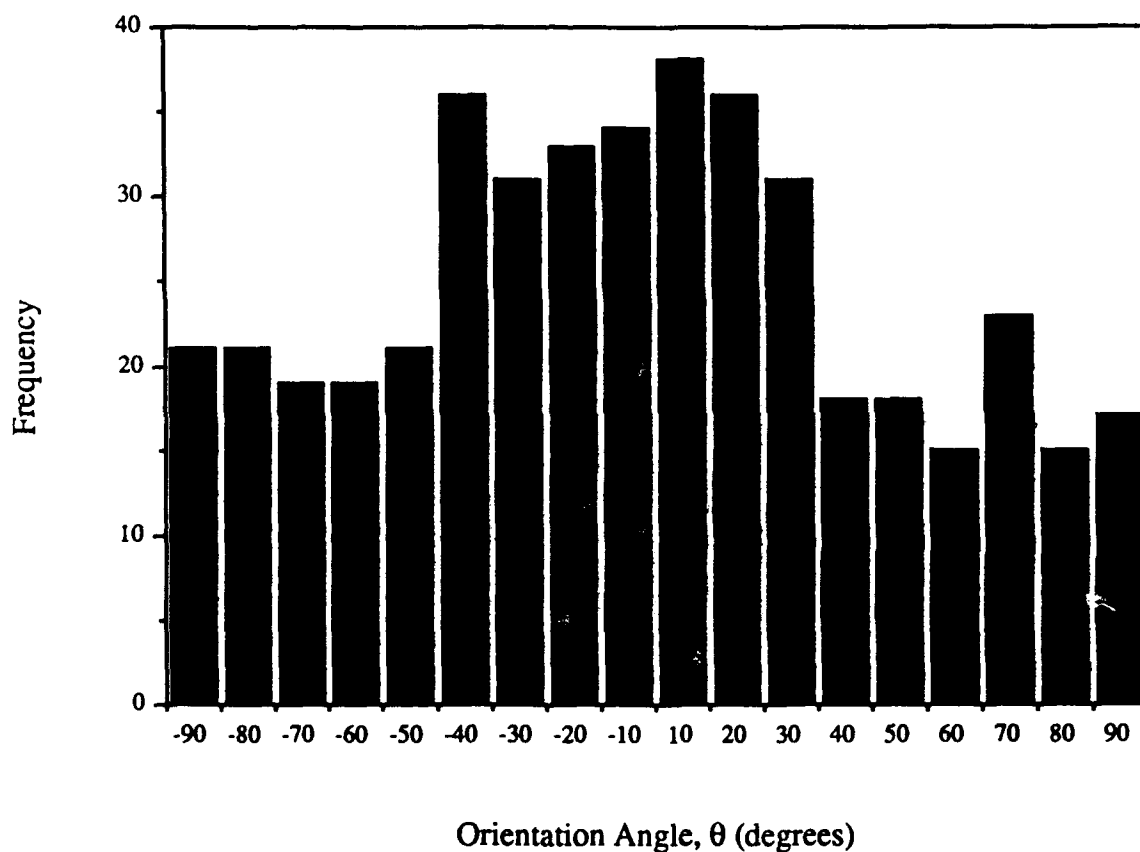


Figure 25. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 18.7\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l, c, e

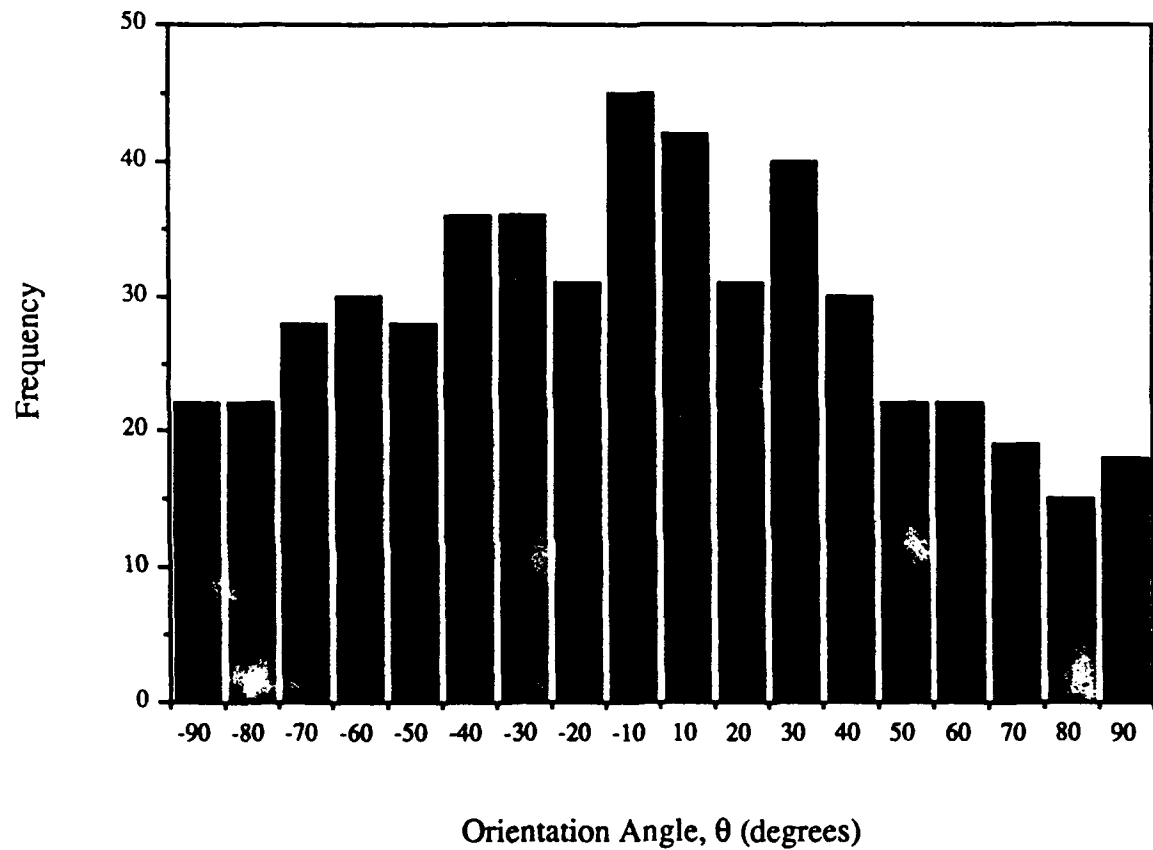


Figure 26. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 37.4\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l, c, e

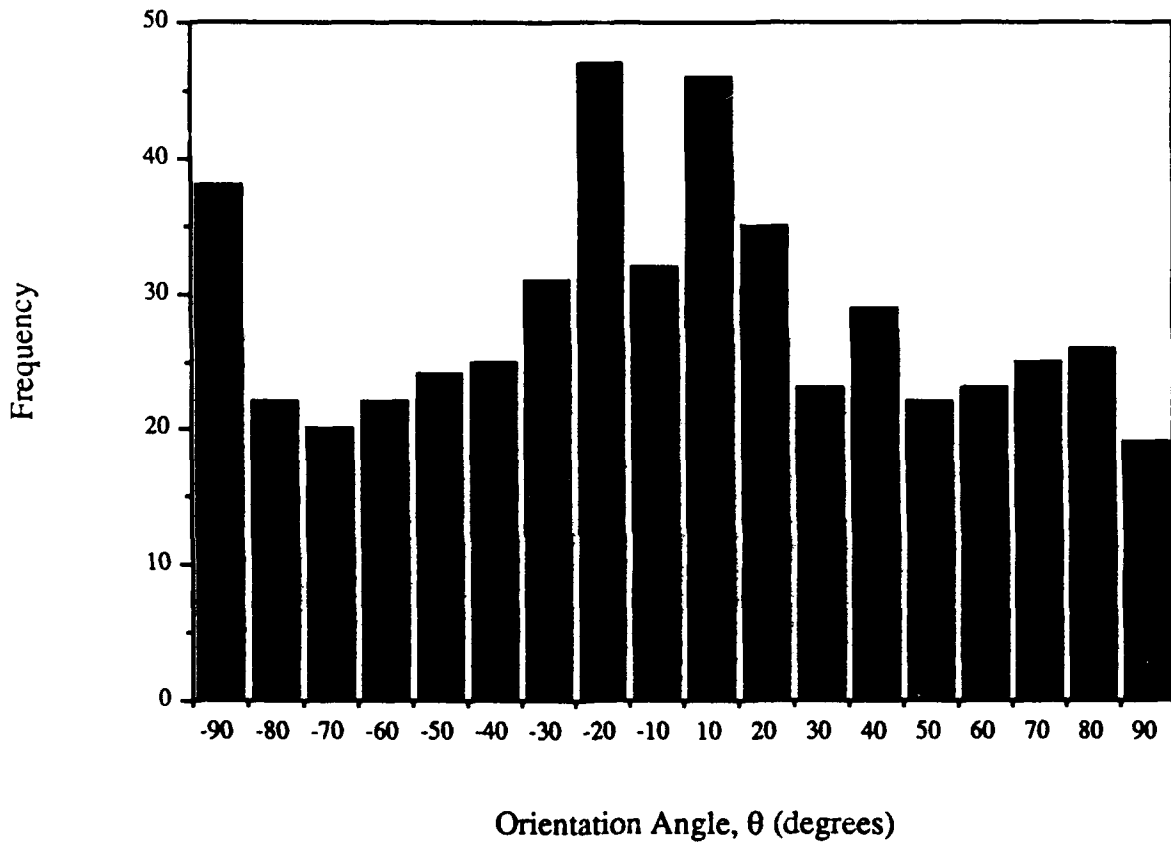


Figure 27. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 56.2\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

Tube Wall Thickness = 2.54 cm; Grid Positions = k, d, l, c, e

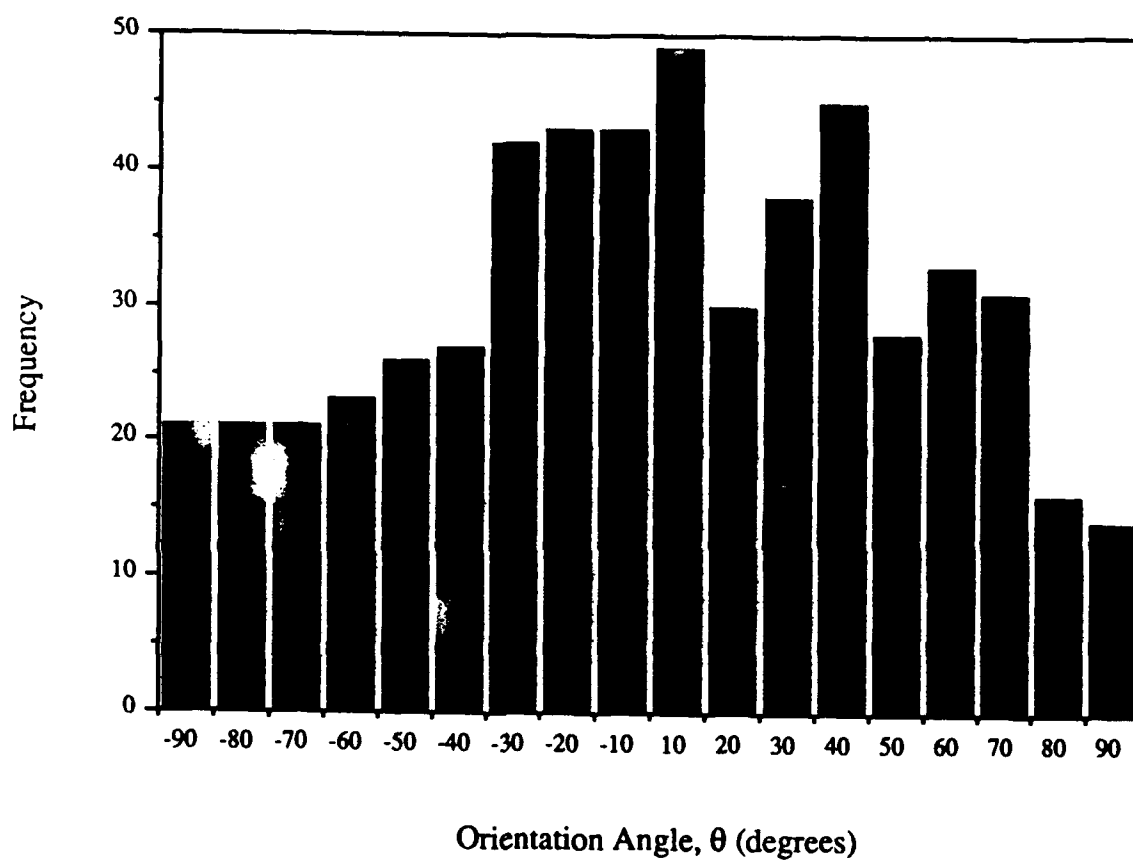


Figure 28. Particle Orientation Histogram for Ottawa 20-30 Sand at $S = 75.0\%$ Compacted to a Dry Density of 1.715 g/cc (107.0 pcf).

TABLE 3. Summary of Particle Orientation Analysis for Particle Orientation Angle, θ (degrees), as a Function of Saturation.

Saturation (%)	k, d, l (a)	c, d, e (b)	k, d, l, c, e (c)
0	-30, -20, +20, +90	-50 to -20	-40, -30, -20
18.7	-50 to -10, +20	-40, -20, +20	-40, +10, +20
37.4	-40, -10, +10	-30, -10, +30	-10, +10, +30
56.2	-20, -10, +10	-90, -20, +10	-90, -20, +10
75.0	-20, +10, +40	-30, +10, +40	-20 to +10, +40

Note: ^aRefer to Figures 14-18 (see Figure 6 also).

^bRefer to Figures 19-23 (see Figure 6 also).

^cRefer to Figures 24-28 (see Figure 6 also).

Results shown are based on the three largest peaks observed in the data.

VI. CONCLUSIONS

Based on the results of this investigation, several conclusions can be drawn with regards to the behavior of compacted unsaturated Ottawa 20-30 sand:

- 1) The compactive energy required to produce a constant dry density specimen by dynamic compaction is strongly dependent on the amount of moisture present. The largest amount of effort is needed to compact specimens at about 20-50% saturation, while the least amount is for dry packing. This can be attributed to variations in overall specimen stiffness with moisture which may be due to the formation of preferred particle orientations and the presence of capillary pressures during compaction.
- 2) For SHPB specimens compacted moist and tested moist, the transmission ratio and wave speed increase from 0% to about 25% saturation, remain constant to about 60% saturation, and then decrease thereafter. An interesting feature in the data is that magnitudes above about 60% saturation are lower than those at 0% saturation. This implies that the presence of moisture during testing may have a lubricating effect as the compressive energy passes through the specimen. Particles would then be able to move somewhat more freely relative to each other and absorb more energy.
- 3) For SHPB specimens compacted moist and tested dry, the transmission ratio and wave speed are generally larger than for specimens compacted moist and tested moist, particularly at about 20% saturation, after which they decrease with increasing saturation. In addition, no plateau region exists and the wave speed remains above that at 0% saturation regardless of the moisture condition during compaction. Therefore, it appears that soil microstructural characteristics developed during compaction which influence the transmission ratio and wave speed, remain intact even after the moisture in the pores has been removed.

- 4) The SHPB data are in general agreement with the findings of Ross (1989). In addition, the results are very similar to those of Hughes and Kelly (1952) who directly measured variations in dilatational velocity with saturation and confining pressure on rock core specimens. This indicates that locked in stresses from compaction in the SHPB tests, and from confining pressure in the rock core tests, may be a significant factor affecting the dynamic response in addition to the moisture condition.
- 5) For the microstructural data analyzed to date, there is some weak dependance of particle orientation on moisture conditions during packing and no particularly strong dependence was evident at any saturation. However, this may be due to the relatively small number of particles examined and the results may change when the remaining data is analyzed. In addition, there may also be some important three-dimensional effects which are not evident in the two-dimensional analysis performed in this study.

VII. RECOMMENDATIONS

The research reported herein has provided some valuable insight into the behavior of unsaturated cohesionless soils. Based on this work, several recommendations can be made for further studies:

- 1) An investigation should be made to study the behavior of Ottawa 20-30 and other sands (such as Eglin and Tyndall sands) at different dry unit weights (ie., variations in compactive energy) to develop correlations among dry density, compactive energy, stress transmission, saturation and microstructural parameters. This would also provide data on the effects of particle size, grain size distribution and particle shape.
- 2) Tests should be conducted to assess the influence of boundary conditions on the static and dynamic properties of unsaturated sand. This would require the

fabrication of an appropriate cell-type device capable of applying controlled confining pressures to a compacted specimen of unsaturated sand that could be used in both the SHPB and a pseudo-static loading system such as the MTS. Some general information about such behavior could also be obtained by performing tests on unsaturated specimens compacted in containers having different wall thicknesses to simulate differing degrees of confinement.

- 3) Some pulse dilatational wave velocity measurements should be made on compacted unsaturated sands at various confining pressures in a modified triaxial testing apparatus. This will provide useful non-destructive data about the influence of pressure on dynamic response that could be correlated with high intensity transient SHPB data and shear strength data. In addition, the triaxial device could also be used to apply anisotropic confinement which would more closely simulate in situ conditions.
- 4) Microstructural studies should be continued and expanded to investigate the influence of variations in compactive energy, saturation, boundary conditions and material type on the static and dynamic behavior of soils. In addition, studies should be initiated to examine three-dimensional spatial orientations in the compacted grain matrix structure. This may yield useful results since some features are inevitably lost when viewing two-dimensional sections.

VIII. REFERENCES

1. ASTM (1990) Annual Book of ASTM Standards - Volume 04.08 Soil and Rock; Dimension Stone; Geosynthetics, American Society for Testing and Materials, Philadelphia, PA, pp. 160-164.
2. Brewer, R. (1964) Fabric and Mineral Analysis of Soils, Wiley and Sons, New York, NY.
3. Campbell, D.A.. (1985) "Sand Fabric as an Indicator of Stress-Strain Response and Enhanced Techniques for its Measurement." Thesis Submitted to the Graduate School at the University of Colorado in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy, Department of Civil, Environmental and Architectural Engineering, 415 p.
4. Charlie, W. A. and Pierce, S. J. (1988) "High Intensity Stress Wave Propagation in Unsaturated Sands." Final Report to US AFOSR/UES Under Contract Number F49620-87-0004, Engineering and Services Center, AFESC/RDCM, Tyndall AFB, FL, September, 20 p.
5. Charlie, W. A., Ross, C. A. and Pierce, S. J. (1990) "Split-Hopkinson Pressure Bar Testing of Unsaturated Sand." Geotechnical Testing Journal, ASTM, Philadelphia, PA, Vol. 13, No. 4, December, pp. 392-300.
6. Charlie, W. A. and Walsh, A. J. (1990) "Centrifuge Modeling of Explosive Induced Stress Waves in Unsaturated Sand." Final Report to AFOSR/UES Under Contract Number F49620-88-C-0053, Air Force Engineering and Services Center, AFESC/RDCM, Tyndall AFB, FL, September, 20 p.
7. Corey, A. T. (1977) Mechanics of Heterogeneous Media, Water Resources Publication, Ft. Collins, CO, 259 p.
8. Crawford, R. E., Higgins, C.J. and Bultmann, E.H. (1974) The Air Force Manual for the Design and Analysis of Hardened Structures, AFWL-TR-74-102, Air Force Weapons Laboratory, Kirtland AFB, Albuquerque, NM, 1118 p.
9. Felice, C. W., Gaffney, E. S., Brown, J. A. and Olsen, J. M. (1987) "Dynamic High Stress Experiments on Soil." Geotechnical Testing Journal, ASTM, Philadelphia, PA, Vol. 10, No. 4, December, pp. 192-202.
10. Fredlund, D. G. (1985) "Soil Mechanics Principles that Embrace Unsaturated Soils." Proceedings of the 11th International Conference on Soil Mechanics and Foundation Engineering, ISSMFE, San Francisco, pp. 465-472.
11. Hughes, D.S. and Kelly, J.L. (1952) "Variation of Elastic Wave Velocity with Saturation in Sandstone." Geophysics, Vol. 17, pp. 739-753.
12. Juang, C. H. and Holtz, R. D. (1986) "Fabric, Pore Size Distribution, and Permeability of Sandy Soils." Journal of the Geotechnical Engineering Division, ASCE, Vol. 112, No GT9, September, pp. 855-868.

13. LaFeber, D. (1965) "The Graphical Presentation of Planar Pore Patterns in Soils." Australian Journal of Soil Research, Vol. 3, August, Australia, pp. 143-164.
14. LaFeber, D. (1972) "Micromorphometric Techniques in Engineering Soil Fabric Analysis." 3rd International Working-Meeting on Soil Micromorphology, September, Wroclaw, Poland, pp. 669-687.
15. Mahmood, A. and Mitchell, J. K. (1974) "Fabric-Property Relationships for Fine-Grained Materials." Clays and Clay Minerals, Vol. 22, pp 397-408.
16. Mitchell, J.K., Chatoian, J.M. and Carpenter, C. C. (1976) "The Influences of Sand Fabric on Liquefaction Behavior." Contract Report S-76-5. US Army Corps of Engineers Waterways Experiment Station, Soils and Pavements Laboratory, Vicksburg, MS, June, 38 p.
17. Mitchell, J.K., Guzikowski, F.J. and Villet, W.C.B. (1978) Fabric Analysis of Undisturbed Sands from Niigata, Japan." Technical Report S-78-11. US Army Corps of Engineers Waterways Experiment Station, Geotechnical Laboratory, Vicksburg, MS, September, 41 p.
18. Mulillis, J. P., Seed, H. B., Chan, C. K., Mitchell, J. K. and Arulanandan, K. (1977) "Effects of Sample Preparation on Sand Liquefaction." Journal of the Geotechnical Engineering Division, ASCE, Vol. 103, No GT2, February, pp. 91-108.
19. Nimmo, J. R. and Akstin, K. C. (1988) "Hydraulic Conductivity of a Sandy Soil at Low Water Content After Compaction by Various Methods." Journal of the Soil Science Society of America, Division S-1-Soil Physics, Vol. 52, No. 2, March/April, pp. 303-310.
20. Oda, M (1972a) "Initial Fabrics and Their Relations to Mechanical Properties of Granular Materials." Soils and Foundations, Japanese Society of Soil Mechanics and Foundation Engineering, Vol. 12, No. 1, March, Tokyo, pp. 17-36.
21. Oda, M (1972b) "The Mechanism of Fabric Changes During Compressional Deformation of Sand." Soils and Foundations, Japanese Society of Soil Mechanics and Foundation Engineering, Vol. 12, No. 2, June, Tokyo, pp. 1-18.
22. Pettijohn, F. J. (1957) Sedimentary Rocks. Harper and Row, New York, NY.
23. Ross, C. A. (1989) "Split-Hopkinson Pressure Bar Tests." Final Report No. ESL-TR-88-2, HQ AFESC/RDCM, Air Force Engineering and Services Center, Tyndall AFB, FL, 80 p.
24. Ross, C. A., Nash, P. T. and Friesenhahn, C. J. (1986) "Pressure Waves in Soils Using a Split-Hopkinson Pressure Bar." Technical Report No. ESL-TR-86-29, USAF Engineering and Services Center, Tyndall AFB, FL, July, 83 p.
25. Turner, F. J. and Weiss, L. E. (1963) Structural Analysis of Metamorphic Tectonites. McGraw-Hill, New York, NY.

26. Veyera, G. E. (1989) "Static and Dynamic Behavior of Compacted Unsaturated Sands." Final Report to US AFOSR/UES Under Contract Number F49620-87-0004, Engineering and Services Center, AFESC/RDCM, Tyndall AFB, FL, September, 20 p.
27. Veyera, G. E. and Fitzpatrick, B. J. (1990) "A Specimen Preparation Technique for Microstructural Analysis of Unsaturated Soil." Final Report to AFOSR/UES Under Contract Number F49620-88-C-0053, Air Force Engineering and Services Center, AFESC/RDCM, Tyndall AFB, FL, September, 20 p.
28. WES (1984) Fundamentals of Protective Design for Conventional Weapons. Design Guide, Chapter 5, US Army Corps of Engineers, Waterways Experiment Station, Soils and Pavements Laboratory, Vicksburg, MS, 333 p.
29. Wu, S., Gray, D. H. and Richart, F. E. (1984) "Capillary Effects on Dynamic Modulus of Sands and Silts." Journal of the Geotechnical Engineering Division, ASCE, Vol. 110, No GT9, September, pp. 1188-1202.
30. Wyllie, M. R., Gregory, A. R. and Gardner, L. W. (1956) "Elastic Wave Velocities in Heterogeneous and Porous Media." Geophysics, Vol. 21, No. 7, January, pp. 41-70.

Report # 31
760-0MG-008
Prof. Hermann Donnert
Report Not Publishable

Report # 32
210-10MG-011
Prof. Robert Granger
Reference 'AIAA 91-0745'
Flow Induced Vibrations of Thin Leading Edges

Report # 33
760-OMG-107
Prof. Ronald Sega
Report Not Publishable

1989 USASF-UES 1989-1990 RESEARCH INITIATION PROGRAM

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the
Universal Energy Systems, Inc.

FINAL REPORT

USE OF NITRONIUM TRIFLATE FOR NITRATION OF NITROGEN HETEROCYCLES

Prepared by:	Clay M. Sharts, Ph.D.
Academic Rank:	Professor
Department and University	Chemistry San Diego State University San Diego, CA 92182-0328
Research Location:	Chemistry Department San Diego State University San Diego, CA 92182-0328
USAF Researcher:	Major Scott Shackelford, USAF USAF FJSRL/NC (AFSC) United State Air Force Academy Colorado Springs, CO 80840-6528
Date	3 July 1991
Contract No:	P. O. S-210-10MG-072

USE OF NITRONIUM TRIFLATE FOR NITRATION OF NITROGEN HETEROCYCLES

by

Clay M. Sharts

ABSTRACT

The reaction of anhydrous lithium nitrate and trifluoromethanesulfonic anhydride (triflic anhydride) in refluxing nitromethane solvent was found to be the most convenient method for preparing nitronium trifluoromethylsulfonate (nitronium triflate). Nitronium triflate in anhydrous nitromethane was found to be a convenient nitrating agent for a limited number of organic compounds. Bromobenzene was converted in 75% yield into ortho- and para-bromonitrobenzene. The objective of the research was to N-nitrate nitrogen heterocycles. Nitration was successful only with 2-pyrrolidone. N-Nitro-2-pyrrolidone was obtained in 46% yield.

ACKNOWLEDGEMENTS

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsoring this research and the Universal Energy Systems for the prompt and timely logistical support provided.

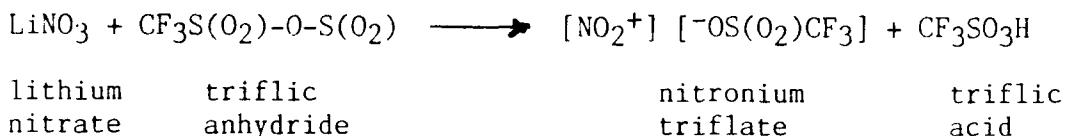
I give special thanks to Major Scott Shackelford for his efforts that made possible my Summer at F. J. Seiler Research Laboratory. He is a true friend and colleague. It has been my pleasure to work with him.

I. INTRODUCTION: It has been recognized for many years that many conventional World War II explosives are too sensitive for use in the high technology environment of modern ships and aircraft. New insensitive high explosives are needed for use in the costly delivery systems used in modern warfare. Classical nitration methods are too vigorous to be used for nitration of many potentially interesting heterocyclic ring systems. The need for new methods of nitration was recognized in an authoritative review by the Explosives Technology Group at the Los Alamos National Laboratory [1]. A review of newer nitration techniques was presented by this investigator in an earlier report written for the 1989-1990 USASF-UES Summer Faculty Research Program [2]. In reference 2 several other references are presented which describe the work upon which the earlier research and the present research was based. For the convenience of the reader the references cited in reference 2 are cited again in this report [3-12].

In earlier scouting work an unreliable and erratic method was developed for in situ synthesis of nitronium triflate (nitronium trifluoromethanesulfonate) from lithium nitrate and triflic anhydride (trifluoromethanesulfonic acid). Nitronium triflate was found to C-nitrate bromobenzene and to N-nitrate 2-pyrrolidone but these reactions were never developed to the point of reliability and reproducibility. The purpose of the present work was to develop the earlier work and apply it to nitration of nitrogen heterocyclic compounds. The major goals of the research conducted for this report were:

1. To achieve one-flask nitrations of heterocyclic compounds using nitronium triflate formed in situ in a low-boiling solvent of low-polarity such as dichloromethane. If this was not possible then:
2. To achieve one-flask nitration of heterocyclic compounds using nitronium triflate formed in situ in a polar solvent such as nitromethane.
3. To form nitronium triflate from metal nitrate salts other than lithium or ammonium nitrates (calcium nitrate in particular).

II. SYNTHESIS OF NITRONIUM TRIFLUOROMETHANESULFONATE: Initially the synthesis of nitronium trifluoromethanesulfonate, trivially named nitronium triflate, was carried out in dichloromethane solvent by reacting dry pulverized lithium nitrate with trifluoromethanesulfonic anhydride (triflic anhydride). It was observed that nitronium triflate formed in refluxing dichloromethane over an extended period of twelve hours to five days. The rates of reaction and yields of product were erratic.



When nitronium triflate was successfully synthesized, it could be determined by visual observation. Solid nitronium triflate sublimed onto a cold finger above dichloromethane or precipitated on the side of the flask at the air dichloromethane interface. In the earlier work [2] nitronium triflate was not isolated but was reacted in the initial reaction flask by adding a dichloromethane solution of 2-pyrrolidone or bromobenzene to the initial flask. In two experiments 2-pyrrolidone was nitrated to form N-nitropyrrolidone in 30-35% yield, but in most experiments other products dominated. The synthesis of nitronium triflate was further demonstrated by nitration of bromobenzene to form a mixture of ortho- and para-bromonitrobenzene.

The unpredictable and erratic formation of nitronium triflate in dichloromethane forced abandonment of this preferred solvent in favor of higher boiling nitromethane and its very high dielectric constant and accompanying problems with separations in the workup. A consistent and reliable formation of nitronium triflate was achieved by reacting lithium nitrate and triflic anhydride for one hour in nitro methane. The synthesis of nitronium triflate was confirmed adding bromobenzene and isolating the nitration products.

Experimental Procedure for Reaction of Lithium Nitrate with Triflic Anhydride. Lithium nitrate was pulverized with a mortar and pestle dried in an oven at 185-190° for a minimum of one day. Dried lithium nitrate and flamed-dried glassware were stored in an oven at greater than 100°. When glassware was assembled, a stream of dry argon was used to maintain dryness. A 100-ml, 3-necked flask was fitted with a compensating dropping funnel, reflux condenser, gas inlet tube and a magnetic stirring bar. Before assembling the apparatus a stoichiometric amount (e.g., 1.13 g, 16.3 mmol) of dried lithium nitrate was added to the hot dry reaction flask. After assembly of the apparatus one stoichiometric amount of triflic anhydride (e.g., 2.2 ml, 13.1 mmol) in 10 ml of dry nitromethane was added. The anhydrous mixture was heated one hour under reflux and cooled over a two-hour period while under an argon atmosphere. At the beginning of the reaction all white solid (lithium nitrate, sp. gr. 2.35) was at the bottom of the flask under the nitromethane solvent (sp. gr. 1.68). After reaction white crystals floated on the nitromethane surface and adhered to the flask walls at the air/solvent interface. A small amount of white solid (LiNO_3) remained at the bottom of the flask.

III. NITRATION OF BROMOBENZENE AS AN ASSAY FOR NITRONIUM TRIFLATE: The last reaction carried out during the 1989 USASF-UES Summer Faculty Research Program was nitration of bromobenzene with nitronium triflate formed in situ from lithium nitrate and triflic anhydride in nitromethane. A mixture of ortho- (32%) and para-bromonitrobenzene (68%) was separated from excess unreacted bromobenzene and a small amount of bromo-2,4-dinitrobenzene. In this study we have used the nitration of bromobenzene as a convenient assay for the formation of nitronium triflate because the volatile bromobenzene is easily removed at reduced pressure from the much less volatile bromonitrobenzenes which can be conveniently identified. The weight of nonvolatile bromonitrobenzenes can be quickly equated to the amount of nitronium triflate formed. The yield of bromonitrobenzenes were consistently 65-77% based on triflic anhydride.

The chemistry described on the preceding pages is summarized by the following two equations:

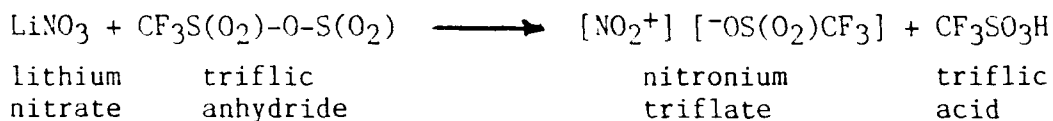


Table I which follows presents a representative summary of nitration reactions of bromobenzene which were used as an assay for formation of nitronium triflate from lithium nitrate and triflic anhydride.

TABLE I. NITRATION OF BROMOBENZENE AS AN ASSAY FOR NITRONIUM TRIFLATE

R u n	Lithium Nitrate g (mmol)	Triflic Anhydride g (mmol)	Solvent NM* or DCM	Time of Reflux hr/days	Bromobenzene g (mmol)	Bromonitrobenzene ortho + para g (mmol) % yield
1.	0.689 (10.0)	2.82 (10.0)	DCM	6 hr	0.157 (10.0)	<0.1 g, nil
2.	0.689 (10.0)	2.82 (10.0)	DCM	18 hr	0.157 (10.0)	0.2 (2.0) 10%
3.	0.689 (10.0)	2.82 (10.0)	DCM	3 days	0.157 (10.0)	0.5 (2.50) 25%
4.	0.689 (10.0)	2.82 (10.0)	DCM	7 days	0.157 (10.0)	0.6 (3.0) 30%
5.	0.689 (10.0)	2.82 (10.0)	NM	6 hr	0.157 (10.0)	1.5 (7.5) 75%
6.	0.689 (10.0)	2.82 (10.0)	NM	1 hr	0.157 (10.0)	1.5 (7.5) 75%
7.	1.38 (20.0)	5.64 (20.0)	NM	1 hr	0.314 (20.0)	3.19 (15.8) 79%

* NM is nitromethane; DCM is dichloromethane.

A literature survey was made of the solubilities of alkali and alkaline earth nitrates in organic solvents in the hope of finding a nitrate more suitable for synthesis of nitronium triflate than lithium nitrate or ammonium nitrate. Table II presents some of the solubility data which was found [8]. Table III is discussed on the next page.

TABLE II. SOLUBILITIES OF GROUP I AND II SALTS IN ANHYDROUS METHANOL

	Grams salt/100 grams anhydrous methanol						
	Fluoride (1.36)	Chloride (1.81)	Bromide (1.95)	Iodide (2.16)	Nitrate	Carbonate	Sulfate
Lithium (0.60)	0.0176	20.98	34.29	...	<u>42.95</u>	0.0555	0.1261
Sodium (0.95)	0.0231	1.401	16.09	62.51	2.936	0.3109	0.0113
Potassium (1.33)	2.286	0.5335	2.080	17.07	0.380	6.165	0.0005
Calcium (0.99)	0.0145	23.26	55.83	67.37	<u>127.13</u>	0.0012	0.0046
Strontium (1.13)	0.0142	18.05	1.061	0.0014	0.0074
Barium (1.35)	0.0044	1.379	0.048	0.0064	0.0063

(Ionic radii in Angstroms are given in parentheses.)

TABLE III. BROMOBENZENE ASSAY FOR NITRONIUM TRIFLATE
FROM CALCIUM NITRATE AND TRIFLIC ANHYDRIDE

R	Calcium Nitrate	Triflic Anhydride	Solvent NM* or DCM	Time of Reflux hr/days	Bromobenzene g (mmol)	Bromonitrobenzene ortho + para g (mmol) % yield
n	g (mmol)	g (mmol)				
1.	1.64 (10.0)	2.82 (10.0)	DCM	5 hr	0.157 (10.0)	0.0 (0.0) 0% No evidence of rxn.
2.	1.64 (10.0)	2.82 (10.0)	DCM	3 days	0.157 (10.0)	0.0 (0.0) 0% No evidence of rxn.
3.	1.64 (10.0)	2.82 (10.0)	NM	12 hr	0.157 (10.0)	0.0 (0.0) 0% No evidence of rxn.
4.	1.64 (10.0)	2.82 (10.0)	NM	3 days	0.157 (10.0)	0.0 (0.0) 0% No evidence of rxn.

* NM is nitromethane; DCM is dichloromethane.

As seen in Table II, calcium nitrate is very soluble in methanol. For this reason we expected it to be soluble in nitromethane and react readily with triflic anhydride to form nitronium triflate. Calcium nitrate did not react with triflic anhydride in either dichloromethane or nitromethane. The results of four attempted reactions are summarized in Table III presented on the previous page.

IV. ONE-FLASK NITRONIUM TRIFLATE NITRATION OF 2-PYRROLIDONE TO FORM N-NITRO-2-PYRROLIDONE: During scouting research in Summer 1989, 2-pyrrolidone was successfully nitrated by nitronium triflate to give N-nitro-2-pyrrolidone in 27-33%. The nitrations were not consistent and in many attempted one-flask reactions, no N-nitro-2-pyrrolidone was obtained. The inconsistency of results was believed to be due to inconsistent formation of the nitrating agent, nitronium triflate.

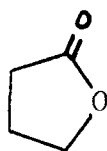
In the work for this report, N-nitro-2-pyrrolidone was obtained in consistent 35-46% yields where nitronium triflate was prepared in nitromethane as described in Section II of this report. Major Shackelford [13] suggested that the maximum yield of N-nitro-2-pyrrolidone may only be 50% because of concomitant O-nitration. In Chart I, structures are shown for the various products that have been formed in the nitronium triflate nitration of 2-pyrrolidone.

Table IV summarizes many of the successful and unsuccessful attempts to prepare N-nitro-2-pyrrolidone. The highest yield of crude N-nitro-2-pyrrolidone based on triflic anhydride was 55% when a 100% excess of 2-pyrrolidone was used. Analysis of crude product by nmr spectroscopy resulted in an estimate of 90% purity which gives a yield of 49% which is less than the 50% figure suggested by Major Shackelford [13].

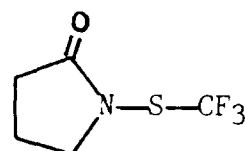
Experimental Procedure for Synthesis of N-Nitro-2-pyrrolidone. Flame-dried glassware was stored in an oven, assembled while hot, and maintained under a dry argon atmosphere. Oven-dried (190°) lithium nitrate (2.03 g, 0.0294 mol) and a 1-cm magnetic stirring bar were

CHART I

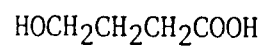
Products from the Nitronium Triflate Nitration of 2-Pyrrolidone



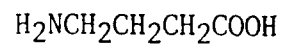
I



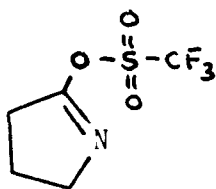
II



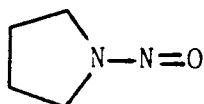
III



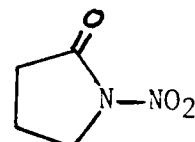
IV



V



VI



VII

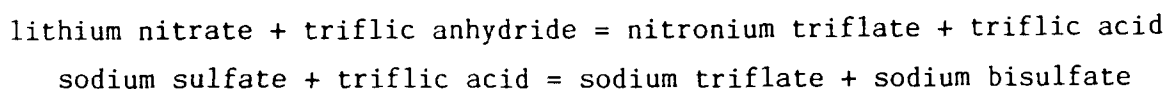
TABLE IV. NITRATION OF 2-PYRROLIDONE WITH NITRONIUM TRIFLATE
TO FORM N-NITRO-2-PYRROLIDONE (NN2P)

<u>R</u>	<u>Lithium Nitrate</u> <u>g (mmol)</u>	<u>Triflic Anhydride</u> <u>g (mmol)</u>	<u>Solvent</u> <u>DCM or NM</u>	<u>Condi-</u> <u>tions</u> <u>t, T</u>	<u>2-Pyrrol-</u> <u>idone</u> <u>g (mmol)</u>	<u>Na₂SO₄</u> <u>g (mmol)</u>	<u>Product Data</u> <u>refer to Chart I</u> <u>misc information</u>
1.	2.85 (41.3)	5.06 (20.2)	DCM 50 ml	24 hr reflux	1.70 (20.0)	5.70 (40.4)	84% I, 16% II no NN2P see CHART I
2.	10.0 (145)	5.72 (20.3)	none	24 hr reflux	1.73 (20.3)	none	18% 2P, 52% I no NN2P
3.	3.0 (43.5)	4.98 (17.7)	none	14 hr reflux	1.72 (20.2)	none	0.134 g, 5% NN2P 95% pure NN2P
4.	none	5.77 (20.5)	DCM 50 ml	2P added	1.70 (20.0)	6.0 (42)	I-V from Chart I
5.	2.75 (40.0)	5.64 (20.0)	none	14 hr reflux	1.70 (20.0)	5.68 (40)	0.80 g (6.0 mmol) 31% NN2P, trace VI
6.	2.75 (40.0)	5.64 (20.0)	NM 1.02 g	4 hr room	1.70 (20.0)	none	0.20 g (1.6 mmol) 7% NN2P, trace VI
7.	1.00 (14.6)	3.69 (13.1)	NM 10 ml	6 hr reflux	0.57 (6.7)	1.89 (13.3)	0.13 g NN2P by ir, nmr
8.	1.90 (27.5)	2.60 (15.6)	NM 15 ml	2 hr 95°C	2.69 (31.0)	4.40 (31.0)	1.14 g crude NN2P 0.91 g pure, 45%
9.	2.03 (29.4)	4.20 (14.9)	NM 10 ml	1 hr 95°C	2.58 (29.8)	4.23 (29.8)	0.93 g NN2P mp = 53-55°C

Notes on Columns of Table IV

Run: Arbitrary #; do not include all experiments.
 Lithium Nitrate: Anhydrous from Baker's LiNO₃ trihydrate.
 Triflic Anhydride: Aldrich best grade.
 Solvent: DCM = dichloromethane; NM = nitromethane.
 Na₂SO₄ Used as base to remove triflic acid.
 Chart I Structures of products given on page 10.
 2P and NN2P 2-Pyrrolidone and N-Nitro-2-Pyrrolidone.

added to a three-necked 100-ml round-bottomed flask which was equipped with a cold finger, condenser and inlet/outlet tubes for maintaining an argon atmosphere. Nitromethane (10 ml) was added to a dry compensating dropping funnel which was placed at the top of the vertical condenser. A dry syringe was used to add 2.60 ml of triflic anhydride (Aldrich catalog 15,853-4, $d = 1.696$, 4.40 g, 15.6 mmol) to the nitromethane in the compensating dropping funnel. The nitromethane/triflic anhydride solution was added rapidly to the dry lithium nitrate and, magnetic stirring was started. The syringe, dropping funnel, and condenser were washed with 5 ml of nitromethane into the reaction flask. The flask was heated at 95°C by an oil bath for 1.5 hours and then permitted to cool to room temperature. Dry sodium sulfate was added (4.40 g, 31.0 mmol) and the reaction mixture stirred for one additional hour in order to neutralize triflic acid formed concomitantly with nitronium triflate as illustrated by the following word equations:



Finally, 2.40 ml of 2-pyrrolidone (Aldrich catalog 24,033-8, $d = 1.120$ 2.69 g, 31.6 mmol) in 5 ml of nitromethane was added dropwise to the stirred reaction solution over a 30-minute period. The reaction mixture became yellow. Stirring of the slurry was maintained with difficulty until the reaction mixture was warmed to 30-35°C. In some runs it was necessary to add an additional 3-5 ml of nitromethane to assure continuous stirring. After 24 hours stirring was stopped and solids separated by vacuum filtration and washed with 2-3 ml of nitromethane. Nitromethane filtrates were combined and added to a separatory funnel where they were extracted three times with 20-ml portions of water.

The water extracts were combined and washed three times with 20-ml portions of dichloromethane. Each of the dichloromethane extracts were evaporated. The first two dichloromethane extracts gave 1.4 g of 2-pyrrolidone which was identified by its ir and nmr spectra. The residue from the third dichloromethane extract was less than 0.05 g.

The nitromethane filtrate was dried over magnesium sulfate and the solvent distilled under reduced pressure and the pumped on at 1 mm to remove residual solvent. The residue amounted to 0.93 g (7.1 mmol) which is a 46% yield based on the amount of triflic anhydride, the most expensive reagent, used. The theoretical yield of N-nitro-2-pyrrolidone is 4.11 g based on 2.69 g of 2-pyrrolidone used initially. Because 1.4 g of 2-pyrrolidone was recovered, the net 2-pyrrolidone was 1.3 g which gives a theoretical yield of 1.98 g. The percentage yield of N-nitro-2-pyrrolidone becomes 47% based on consumed 2-pyrrolidone which is remarkably close to 46% based on triflic anhydride and completely consistent with Major Shackelford's suggestion [13] which requires two moles of triflic anhydride per mole of 2-pyrrolidone rather than the 1:1 ratio that we had assumed initially.

V. ONE-FLASK NITRONIUM TRIFLATE NITRATION OF 2-PYRROLIDONE TO FORM N-NITRO-2-PYRROLIDONE: After the formation of nitronium triflate from lithium nitrate and triflic anhydride in nitromethane was confirmed by the successful nitration of bromobenzene, the nitration of 2-imidizolidone was attempted. In Table V the attempts to nitrate 2-imidizolidone with nitronium triflate are summarized. Table V gives the amounts of reagents and a brief description of the results for those experiments which were carried through the workup procedure. A variety of workup procedures were tried but none succeeded in isolating nitrated compounds. Imidizolidone nitration experiments which were abandoned are not included. The solvent for all experiments was nitromethane.

Infrared and nmr spectra and melting points were observed for all solid products isolated in the experiments reported in Table V. A flame test on a platinum wire was used to identify nonmelting or high-melting solids as lithium salts. Lithium salts give an easily identified red flame which clearly differentiates them from sodium or potassium salts. Most of the products isolated were lithium salts.

TABLE V. NITRATION OF 2-IMIDIZOLIDONE WITH NITRONIUM TRIFLATE

Run, NB ID	Lithium Nitrate g, (mmol)	Mls NM Solv	Triflic Anhydride g (mmol)	Reaction Conditions time, temp	2-Imidizol- idone g (mmol)	Comments and other information
1. I-44	0.445 (5.11)	7.5	0.86 ml 1.44(5.11)	4 hr, 105°C NT crystals	0.522 (6.06)	14 hr, rm temp 0.1 g benzoic NaOH workup
2. I-50	0.984 (14.1)	15	2.00 ml 3.35(11.9)	4 hr, 105°C NT crystals	0.594 (6.90)	0.1 g benzoic mp 120-121 NaOH workup
3. I-56	0.987 (14.3)	10	1.70 ml 2.85(10.1)	4 hr, 105°C NT crystals	0.457 (5.30)	0.1 g benzoic NaOH workup
4. I-60	1.828 (26.5)	10	3.70 ml 6.20(22.0)	4 hr reflux NT crystals	0.947 (11.0)	DCM extraction Only salts +Li flame test
5. I-65	0.918 (13.3)	10	1.90 ml 3.18(11.3)	4 hr reflux clear soln	0.885 (10.2)	CF ₃ SO ₂ in ir no NO ₂ in ir +Li flame test
6. I-68	0.887 (12.9)	12	1.80 ml 3.02(10.7)	2.5 hr reflux NT crystals	0.850 (9.9)	only salts, ir red flame
7. I-70	1.821 (26.5)	12	2.60 4.36(15.5)	2.0 hr reflux NT crystals	0.812 (9.4)	benzoic by ir NaOH workup

Notes on Columns of Table V

Run/NB ID: Arbitrary number and notebook page reference.

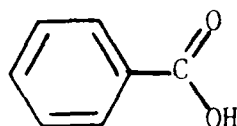
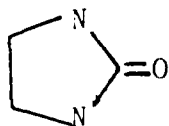
Lithium Nitrate: Anhydrous from LiNO₃·3H₂O; grams, millimoles.

Mls NM Solv: Milliliters of Nitromethane solvent used.

Triflic Anhydride: Aldrich, 15,853-4; milliliters, grams, millimoles.

Rxn. Conditions: for forming Nitronium Triflate (NT);
NT Crystals means that NT observed at NM surface.

2-Imidizolidone Grams, millimoles of 2-I added after NT formed.



It is difficult to see how 2-imidizolidone can be converted to benzoic acid. It was observed in those experiments in which sodium hydroxide was used in the workup and also in one experiment in which a neutral

workup was used. The experimental procedure for one of the experiments is described in the next paragraph. In the run described there was an excess of nitronium triflate. The target of the reaction was the known compound N,N'-dinitro-2-imidazolidone, with melting point 216-217°C.

Run 5, Table V; Nitration of 2-Imidizolidone. In a dry 100-ml flask under a nitrogen atmosphere were mixed 0.918 g (13.3 mmol) of lithium nitrate, 1.90 ml (3.18 g, 11.3 mmol) of triflic anhydride, and 10 ml of nitromethane solvent. The mixture was heated and magnetically stirred under reflux for four hours. During the four hours crystals of lower density than nitromethane formed at the air-nitromethane interface and the flask side above the interface. Lithium nitrate remained at the bottom until reacted. After nitronium triflate was formed, 0.885 g of 2-imidizolidone (10.2 mmol) was added and the flask agitated to achieve mixing. All crystals dissolved within 10 minutes except for the small amount of excess lithium nitrate in the bottom of the flask. Dry sodium sulfate (3.3 g, 23.2 mmol) was added to neutralize triflic acid. Solids were then separated by filtration.

The nitromethane solution was extracted with three 10-ml portions of water. The nitromethane solution was then dried over 0.5 g MgSO_4 and the MgSO_4 separated by filtration. Nitromethane solvent was evaporated under reduced pressure with a rotary evaporator. A yellow oil/crystal mass remained which amounted to less than 0.10 g. A few white crystals (designated I-65A) were on the upper portion of the evaporating flask. The infrared spectrum of U-I-65A did not contain absorptions characteristic of C-Nitro or N-Nitro groups [very strong absorptions at 1570-1540 or 1630-1530 cm^{-1} ; strong at 1390-1340 or 1315-1260 cm^{-1}]. On the basis of the melting point (121°C) and infrared absorptions at 1604, 1584, 1496, 935, 708 and 685, sample I-65A was identified as benzoic acid. Earlier in another experiment, a similar sample was identified as benzoic acid by mass spectroscopy performed at the Seiler Laboratory at the U. S. Air Force Academy.

The magnesium sulfate used for drying the nitromethane layer was washed with 5 ml of dichloromethane. The water extracts of the nitromethane layer were extracted with three 10-ml portions of dichloromethane. Dichloromethane extracts were combined and solvent removed with a rotary evaporator. The residue (designated I-65B) amounted to less than 0.1 g. Residue I-65B was identified as starting material 2-imidizolidone by its infrared and nmr spectra.

The original water extracts of the reaction solution were evaporated to give more than 3 g of white and yellow crystalline and amorphous solid designated as I-65C1. A portion of I-65C1 was washed with hexane, air dried and labeled I-65C2. Sample I-65C2 melted over a wide range, most melting occurring at 197-210°C. The infrared spectra of I-65C1 and I-65C2 were very similar and had a carbonyl absorption at 1689 cm^{-1} which was probably due to benzoic acid. A broad peak occurred between 3300-3400 cm^{-1} (OH or NH). A very strong absorption at 1292 cm^{-1} was consistent with the trifluoromethyl group. There were no absorptions consistent with a nitration product. We concluded I-65C1 and I-65C2 were essentially the same.

The nmr spectrum of 7 mg of I-65C1 in 0.8 ml deuterium oxide showed absorptions at 3.2, 3.3 and 4.5 ppm. The peak at 4.5 is a water peak which is merged with the N-H protons of 2-imidizolidone. Sample I-65C1 in deuteroacetone had proton absorptions at 3.48 and 3.68 ppm. On addition of authentic 2-imidizolidone to the deuteroacetone solution of I-65C1 there were peaks at 3.48, 3.68 and 4.36 ppm. The major organic substance is recovered 2-imidizolidone.

Additional work was carried out on samples derived from I-65C1. Many infrared and nmr spectra were observed. Melting points were observed and flame tests made. The conclusion was that most of the material in the water layer was lithium salts contaminated by starting material. Certainly one of the products was lithium trifluoromethylsulfonate. At no time could evidence be found for a nitro compound.

Discussion of Reaction of 2-Imidizolidone with Nitronium Triflate.

This discussion assumes that lithium nitrate and triflic anhydride react to form nitronium triflate. It is possible that nitronium triflate does nitrate 2-imidizolidone to form N-nitro-2-imidizolidinone and/or N,N'-dinitro-2-imidizolidone, but if they are formed, they are not stable to the reaction conditions or the conditions of work up. The formation of benzoic acid is evidence of major rearrangement. Suri [3] was able to nitrate 2-imidizolidone using nitronium trifluoroacetate formed from ammonium nitrate and trifluoroacetic anhydride to give N,N'-dinitro-2-imidizolidone. The strongest acid present for Suri was trifluoroacetic acid. We have the very strong triflic acid present which may cause our difficulties.

Beilstein reports that 2-imidizolidone is dinitrated by nitric acid to form N,N'-dinitro-2-pyrrolidone. This dinitro compound undergoes hydrolysis in base to form N,N'-dinitroethylenediamine. We did not see evidence for this compound in our basic workups of reaction mixtures.

Our failures may be due to the possibility of O-nitration as well as N-nitration. Or it may be due to triflic acid present in the in situ nitrations. It is apparent that we must isolate nitronium triflate and use it as a nitrating agent under neutral conditions or under weakly acidic or basic solutions.

VI. ONE-FLASK NITRONIUM TRIFLATE NITRATION OF PYRROLIDINE

Attempts were made to nitrate pyrrolidine using the techniques developed for bromobenzene and 2-pyrrolidone. Suri [3] synthesized N-nitropyrrolidine using nitronium trifluoroacetate to nitrate pyrrolidine. He reported the infrared spectrum of this compound. Four of our experiments are summarized in Table VI. In none of these experiments did we find evidence for a nitro compound.

TABLE V. NITRATION OF 2-IMIDIZOLIDONE WITH NITRONIUM TRIFLATE

Run,	Lithium Nitrate g, (mmol)	Mls NM Solv	Triflic Anhydride g (mmol)	Reaction Conditions time, temp	Pyrrolidine g (mmol)	Comments and other information
1.	2.02 (29.3)	10	4.94 (17.5)	1 hr reflux	1.36 (19.2)	K ₃ PO ₄ as base no products
2.	1.13 (16.4)	10	3.69 (13.1)	1 hr reflux	0.68 (9.6)	K ₃ PO ₄ as base no products
3.	1.57 (22.8)	10	6.03 (21.4)	1 hr reflux	1.53 (21.6)	K ₃ PO ₄ as base no products
4.	2.30 (33.4)	10	5.37 (19.0)	1 hr reflux	1.53 (21.6)	K ₃ PO ₄ as base no products

Notes on Columns of Table V

Run/NB ID: Arbitrary number.

Lithium Nitrate: Anhydrous from LiNO₃·3H₂O; grams, millimoles.

Mls NM Solv: Milliliters of Nitromethane solvent used.

Triflic Anhydride: Aldrich, 15,853-4; milliliters, grams, millimoles.

Rxn. Conditions: for forming Nitronium Triflate (NT).

Pyrrolidine: Grams, millimoles of P added after NT formed.

VII. SIGNIFICANT FINDINGS AND CONCLUSIONS. A new method for synthesis of nitronium triflate has been developed. Lithium nitrate reacts with triflic anhydride in refluxing nitromethane in one hour. We have found that nitronium triflate sublimes and can be isolated but have not used isolated nitronium triflate for nitration reactions. Our objectives have been in situ nitration. We have succeeded only in nitrating bromobenzene and 2-pyrrolidone. Mixed ortho- and para-bromonitrobenzene was obtained in 75% yield; N-nitro-2-pyrrolidone was obtained in 46% yield. No nitro products were obtained from pyrrolidine or 2-imidizolidone. We did not achieve the goal of this study, the in situ nitration of nitrogen heterocyclic compounds. We believe this is due to the presence of an equivalent of triflic acid that accompanies nitronium triflate and the presence of unreacted triflic anhydride.

Nitronium triflate can be isolated by sublimation from the reaction flask in which it is formed. The next investigation should be the use of isolated nitronium triflate in neutral solvents with a base present to pick up triflic acid as it forms.

VIII. RECOMMENDATIONS BASED ON SIGNIFICANT FINDINGS AND CONCLUSIONS.

No further investigations of in situ nitration by nitronium triflate should be made. Instead, an experimental technique for isolating nitronium triflate should be accomplished. Nitronium triflate should then be used in neutral solvents such as dichloromethane in the presence of a base such as sodium sulfate for the nitration of difficultly nitrate organic compounds.

IX. REFERENCES. References follow on the next two pages.

REFERENCES

1. Michael D. Coburn, Betty W. Harris, Klen-Yin Lee, M. M. Stinecipher, and Helen H. Hayden; "Explosives Synthesis at Los Alamos"; I&EC PRODUCT RESEARCH/DEVELOPMENT, 1986, 25, 68-72. Chem. Abstr., 1986, 104, 112304z.
2. Clay M. Sharts; "A Convenient Preparation of Nitronium Triflate and Its Use for Nitration", Final Report for Universal Energy Systems, Inc., Contract Number F49620-88-C-0053 for United State Air Force Office of Scientific Research.
3. Stephen J. Kuhn and George A. Olah; "Aromatic Substitution. VII. Friedel-Crafts Type Nitration of Aromatics"; J. Am. Chem. Soc., 1962, 83, 4564-4571. Chem. Abstr., 1962, 56, 8607g.
4. George A. Olah, Stephen J. Kuhn, and Sylvia H. Flood; "Aromatic Substitution. VIII. Mechanism of the Nitronium Tetrafluoroborate Nitration of Alkylbenzenes in Tetramethylene Sulfone Solution. Remarks on Certain Aspects of Electrophilic Aromatic Substitution"; J. Am. Chem. Soc., 1962, 83, 4571-80. Chem. Abstr., 1962, 56, 8607f.
5. Stephen J. Kuhn and George A. Olah; "Aromatic Substitution. IX. "Nitronium Tetrafluoroborate Nitration of Halobenzenes in Tetramethylene Sulfone Solution"; J. Am. Chem. Soc., 1962, 83, 4581-4585. Chem. Abstr., 1962, 56, 8607h.
6. S. C. Suri and R. D. Chapman; "A Convenient Method for N-Nitration Using Ammonium Nitrate/Trifluoroacetic Anhydride". SYNTHESIS, 1988, 473-476. Chem. Abstr., 1989, 110, 192709x.
7. Clifford L. Coon, William G. Blucher, and Marion E. Hill; "Aromatic Nitration with Nitric Acid and Trifluoromethanesulfonic Acid; J. Org. Chem., 1973, 38, 4243-4248. Chem. Abstr., 1974, 80, 14671h.

REFERENCES (continued)

8. S. A. Shackelford, J. W. Beckmann and J. S. Wilkes; "Deuterium Isotope Effects in the Thermochemical Decomposition of Liquid 2,4,6-Trinitrotoluene: Application to Mechanistic Studies Using Isothermal Differential Scanning Calorimetry Analysis"; J. Org. Chem., 1977, 42, 4201-4206. Chem. Abstr., 1978, 88, 6022x.
9. S. A. Shackelford, M. B. Coolidge, B. B. Goshgarian, B. A. Loving, R. N. Rogers, J. L. Janney, and M. H. Ebinger; "Deuterium Isotope Effects in Condensed-Phase Thermochemical Decomposition Reactions of Octahydro-1,3,5,7-tetrazocine; J. Phys. Chem., 1985, 89, 3118-3126. Chem. Abstr., 1985, 103, 36985j
10. R. E. Harner, J. B. Sydnor, and E. S. Gilreath; "Solubilities of anhydrous Ionic Substances in Absolute Methanol"; J. Chem. Eng. Data., 1963, 8, 411-412. Chem. Abstr., 1963, 59, 8178g.
11. B. S. Krumgal'z, V. A. Smirnova, Yu. I. Gerkzhberg; "Solubility of Lithium Salts in Acetone and Methyl Ethyl Ketone in the -50 to +50° Range"; Zh. Prikl. Khim. (Leningrad), 1973, 46, 237-239. Chem. Abstr., 1973, 68, 102560k.
12. P. Srivastava, Mohd. M. Husain, and Ram Gopal; "Solubilities of Inorganic Salts and Tetraalkylammonium Iodides in Sulfolane at Several Temperatures"; J. Chem. Eng. Data., 1985, 30, 144-145. Chem. Abstr., 1985, 102, 155665u.
13. S. A. Shackelford, F. J. Seiler Research Laboratory, USAF Academy, Colorado; private conversations on this work, February 1991.

Report # 35
760-OMG-053
Prof. Walter Trafton
Report Not Publishable

REPORT SUBMITTED TO AFOSR RESEARCH INITIATION PROGRAM

TITLE: ACTIVE CONTROL OF DYNAMIC STALL PHENOMENA

BY:

T. R. TROUTT

ASSOCIATE PROFESSOR

AND

J. A. ALBERTSON-LOVATO

GRADUATE STUDENT

DEPARTMENT OF MECHANICAL AND MATERIALS ENGINEERING

WASHINGTON STATE UNIVERSITY

PULLMAN, WA 99164-2920

This experimental research program focuses on the active control of dynamic stall phenomena. Previous studies of constantly pitched and oscillating airfoils have concluded that significant enhancements of lift occur prior to vortex manifestation on the airfoil surface. A logical subsequent step is to evaluate means of controlling the flow, delaying vortex formation, and extending enhanced aerodynamic force capabilities. This study focuses upon identifying control parameters necessary to actively force the unsteady flow over a NACA 0015 airfoil pitched at constant rates. The characteristic frequencies of the specific flows are first determined, and then a flow visualization analysis is performed on the unsteady flow field under natural and controlled conditions. Results show that forcing at the subharmonic flow frequencies provides the greatest reduction in the upper surface separation region area and the longest delay in dynamic stall vortex formation. These results indicate that extensions of high lift would be probable under similar active forcing conditions.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support and guidance of Capt Scott J. Schreck and Capt Dave J. Bunker. They also thank Mr. Bobby Hatfield for his invaluable efforts in constructing the experimental models and assisting in apparatus design. The help of SSgt Young Paek is appreciated in electronic data analysis support and construction of the timing circuit. The assistance of Mr. Jim Smith in the assembly of the valve system to supply the tangential pulsed air and the design of the timing circuit was invaluable. The authors would also like to thank the support staff of the Frank J. Seiler Research Laboratory for general assistance in many aspects of this project. The financial support of the Air Force Office of Scientific Research and Universal Energy Systems is also greatly valued.

1: INTRODUCTION

The ability of high-performance aircraft to perform in combat situations is governed by aircraft maneuverability and stability. Design changes in air to air weapons have increased the need for fighter-aircraft to execute increasingly difficult maneuvers, in which the best possible instantaneous turning performance is essential to combat superiority. This specification has lead to an interest in aircraft lift enhancement and stability control.

The control systems on air to air combat weapons have attained all-aspect capabilities. This allows them to move in conjunction with aircraft maneuvers and to couple firing systems with flight control systems. Because of these all-aspect weapons, combat effectiveness has become more dependent upon the unsteady performance of the aircraft, namely the ability to execute increasingly difficult maneuvers in which the best possible turning ratio is essential and to maintain pilot control necessary for weapon aiming accuracy and stall prevention. The need for high turning rates result in flight at high angles of attack, which nominally leads to separation on the wing. This separation can lead to increased drag, wing buffeting, and stability and control problems which degrade combat effectiveness and could lead to loss of aircraft (Whitford, 1987).

The leading edge vortex which forms over an airfoil or wing during a constant or oscillating pitching motion has been shown to be connected to aircraft performance. Figure 1, taken from flow visualization results of Walker and Chou (1987), shows the characteristics of such a vortex. Visually, the dynamic stall vortex begin as a small leading edge separation bubble and grows with angle of attack until it detaches from the upper airfoil surface. This vortex is associated

with large decreases in upper surface pressure values and increased pressure gradients. The presence of the vortex on the upper airfoil or wing surface is preceded by enhanced values of lift and drag. It is necessary to examine the events leading up to vortex formation in order to determine the nature of the flow conditions which create the increased aerodynamic forces. Knowing these conditions, it should be possible to actively control the flow and maintain enhanced force characteristics.

Although only a few active control methods have been specifically developed for the complicated mechanisms of unsteady separation and dynamic stall phenomena, numerous control techniques have been explored in static airfoil cases which hold promise. Significant advancements have been achieved in the area of acoustic control of flow over static airfoils, step flows, and mixing layers and wakes. Employing both internal and external acoustic forcing, researchers have been successful in both reducing the separated region over the airfoil and delaying stall in conjunction with enhanced lift coefficients. Although it is not practical to install an acoustic generator on a aircraft, there is sufficient engine bleed air from most conventional aircraft to supply an alternating air blowing and suction system. It has been shown that the mechanisms of acoustic control are similar to air injection/suction (Williams, 1991). Both spanwise and tangential blowing have been employed successfully to delay leading edge vortex growth on delta wings and thus extend the regime of stable, controlled flow over the upper wing surface.

The previous approaches to static airfoil control can be extended to flow about unsteady airfoils with modifications. The flow over an airfoil pitching at constant rates is comparable to static flow prior to leading edge separation (Carr, 1988). Thus, it is reasonable to assume that the established techniques currently

used to control free and bounded shear flows should be applicable here. Gad-el-Hak (1986) noted that changing the leading edge flow characteristics should lead to changes in the entire flow field. This study focuses on the determination of the characteristic flow frequencies associated with the Kelvin-Helmholtz instabilities for constantly pitched airfoils. This knowledge will then be used to explore the benefits of external acoustic forcing of the flow. The qualitative and quantitative information produced by this study provide important insight on both the mechanisms leading to the lift enhancement characteristic to such flows and also the means by which to manipulate these events to improve airfoil performance.

2: SUBJECT OF INVESTIGATION AND MAJOR OBJECTIVES

Although significantly more energetic and complex, the unsteady flow around a constantly pitched, two-dimensional airfoil in part bears resemblance to free shear and reattaching separated flows such as mixing layers and flows over downstream facing steps. In the step flow, since the step height is large compared with the upstream boundary layer thickness, the flow field immediately downstream of the step can be considered a boundary layer in transition to a free mixing layer (Troutt, Scheelke, and Norman, 1984). Similarly, the flow over a pitching airfoil can be initially associated with a free shear layer, and as such is susceptible to small perturbations via the Kelvin-Helmholtz instabilities (Ho and Huerre, 1984).

A significant amount of work has been accomplished which links the initial Kelvin-Helmholtz instability frequency with active forcing of flow fields. The

ability to apply the successful static studies to the dynamic airfoil flow rests upon two recent experimental determinations:

- 1) The unsteady separated region which occurs prior to the dynamic stall vortex formation is similar to the separation region occurring over a static airfoil and other less physically similar flows (such as the flow over a downstream facing step, etc) (Gad-el-Hak and Mangalam, 1991).
- 2) The lift enhancement associated with dynamic flows embodies itself prior to the development of the dynamic stall vortex, and thus the initial unsteady separation region is what must be focused upon when evaluating active controls (Albertson, Troutt, and Kedzie, 1988).

This research study involves the evaluation and the development of acoustic and tangential pulsed air blowing control methods for a constantly pitched NACA 0015 airfoil. These applied methods are then analyzed using flow visualization. A key step in the utilization of these methods is the accurate determination of the characteristic frequencies associated with the Kelvin-Helmholtz instabilities on both a static and a dynamic airfoil. This is accomplished by visual observation of the instability waves and by the use of hot-film anemometry to observe the associated doppler-like pulse created by the instabilities. This study covers the low range of non-dimensional pitch rates, α^+ from 0.01 to 0.15, since the slower moving airfoil bears an even greater resemblance to the static cases, and this realm is where the greatest difference in aerodynamic force coefficients occurs.

The primary objective of the proposed experiments is thus to determine the characteristic flow frequencies that are coupled with dynamic stall flows and to establish that the associated energetic flows can be effectively controlled by proven active means. Once the key forcing parameters are identified, the actively modified flow behavior can be evaluated and conclusions can be drawn concerning the benefits of such forcing and subsequent evaluations can be made to generalize the results to new designs or techniques for practical applications.

3: SUMMARY OF PREVIOUS WORK

Recent significant advances in air combat capabilities have shifted the focus of fighter aircraft design from performance specifications (energy) to extended maneuverability capabilities. The development of all-aspect short-range infrared guided missiles and all-aspect capability guns coupled with aircraft position controls, the emphasis in close air combat design has shifted to enhanced maneuverability. (Herbst, 1983a & 1983b). When airfoils or wings are pitched rapidly beyond their static-stall angle of attack, (the unsteady motion) results in a delay of stall and aerodynamic force significantly greater than the static stall case. The ensuing complex, unsteady flow process is characterized in general as dynamic stall, and has been the subject of numerous investigations. (McCroskey, 1982, & Carr, 1988)

Prior to embarking on any analysis of the flow events leading up to and including the development of dynamic stall, a comprehensive review of research results to date is necessary. The interest in forced, unsteady, separated flows

evolved from the first studies of dynamic overshoot in lift on helicopter rotors (Ham and Garelick, 1968) to current investigations in two areas, low amplitude oscillating airfoils and wings and high amplitude, high rate constant pitching motions. For the purposes of this experimental study, only constant rate two-dimensional motions are being considered. The reasoning for this focus is specified later.

The interest in unsteady separated flows lies in the large transient aerodynamic forces they generate. The focus of early investigations into two-dimensional, unsteady separation focused upon the inherent presence of the dynamic stall vortex, the leading edge vortical structure which is characteristic of these flows. Such flows are characterized by a non-dimensional pitch rate, α^+ , defined as $\alpha c/U_\infty$, where α is the actual pitch rate, c is the airfoil chord, and U_∞ is the free stream velocity. Cook (1987) showed that, for inertia force dominated flows, the non-dimensional pitch rate could provide effects that are orders of magnitude greater than the Reynolds number effects. Walker and Chou (1987) substantiated this analysis through experimentation.

Walker and Chou (1987) obtained a detailed qualitative sequence of events leading up to dynamic stall through smoke wire flow visualization of the flow. Illustrated by the author in Figure 1, dynamic stall vortex development is shown for an α^+ of 0.1, with the airfoil pivot point located at 0.25c. 1(a): The flow transitions from quasi-steady flow with symmetric alternating vortices shed into the wake to an asymmetric, unsteady flow. 1(b): The upper surface separation zone moves forward from the trailing edge, accompanied by an increase in relative strength of the trailing edge counter-clockwise vortices. 1(c): As the separation zone grows, a shear layer interaction between the inviscid outer flow and the viscous boundary layer produces a recirculation region where a number of small clockwise vortices

appear. These clockwise vortices are similar to structures caused by Kelvin-Helmholtz instabilities in free shear flows. 1(d): The shear vortices are absorbed into a turbulent region which has formed and moved forward from the trailing edge. A leading edge separation bubble has formed near the leading edge. 1(e): The leading edge bubble manifests itself in the form of a clockwise vortex, the dynamic stall vortex, and grows on a scale comparable to the airfoil chord. 1(f): The dynamic stall vortex has detached from the airfoil surface and convects downstream perpendicular to the airfoil chord. A trailing edge vortex forms on the trailing edge.

The development of the unsteady separation region over a constantly pitched airfoil and the growth of the dynamic stall vortex results in extensive variations in the upper surface pressure field. Francis and Keesee (1985) conducted an analysis of the pressure field on a constantly pitched airfoil. They found that, although there is significant dependence on non-dimensional pitch rate, the qualitative features of the surface pressure field are similar in various dynamic stall cases. At attack angles lower than the corresponding static stall values, the configuration of the pressure distributions resembles quasisteady (static) flow, except for a lag in pressure coefficient magnitudes. As the angle of attack passes the static stall point, the pressure distribution for an attached flow persists, again with pressure coefficient magnitudes in excess of steady flow values. The suction peak detaches from the leading edge region and moves downstream along the suction surface as the angle of attack continues to increase. Comparing the behavior of the upper surface pressure field with flow visualization data by Walker, Helin, and Strickland (1985) reveals that the suction peak position corresponds to vortex position over the airfoil. This

was confirmed by Albertson, et al (1987) in a digital image analysis study tracking vortex movement over the airfoil.

The manifestation of these increased suction magnitudes for unsteady airfoil flow results in accentuated lift and drag values as compared to static cases. Lift coefficients more than double the static stall values have been determined for non-dimensional pitch rates as low as 0.03 (Albertson, Troutt, and Kedzie, 1988). Drag coefficient values were found to increase correspondingly, although analysis of the lift to drag ratios showed airfoil performance enhancements well above static airfoil experiments.

The non-dimensional pitch rate, α^+ , has been identified numerous times as a characteristic descriptive parameter for dynamic stall flows. The variation in the flow field with respect to α^+ can be categorized into three areas, flow visualization, surface pressure measurements, and surface velocity field measurements. In a comprehensive flow visualization analysis, Walker, Helin, and Strickland (1985) found that increasing α^+ from 0.2 to 0.6 both delayed flow separation and dynamic stall vortex formation. In addition, at the higher α^+ value, significant secondary vortical structures appear. Albertson, et al (1987) used digital image analysis to quantify the growth and movement of the leading edge vortex with respect to angle of attack. They determined that, for both α^+ values of 0.1 and 0.2, the dynamic stall vortex experiences a short period of slow growth, suggesting a quasi-stable period, followed by rapid growth and eventual vortex detachment. The angle of attack at which the vortex commences enhanced growth coincides closely to the point when dynamic stall occurs. It was also ascertained that the commencement of rapid vortex growth coincides with the attack angle at which the vortex center is located approximately over the quarter chord position. Jumper, Schreck, and

Dimmick (1985) suggested that a plateauing in their lift curves with respect to angle of attack corresponded with the quarter chord separation and the subsequent dynamic stall vortex detachment from the airfoil surface. These results indicate that there indeed is a connection between the quarter chord separation point and the overall aerodynamic performance of the airfoil which bears further investigation.

The effect of non-dimensional pitch rate change on the upper surface pressure field has been the subject of several pressure surface studies. Experimental analyses performed by Walker, Helin, and Chou (1985) indicated that lift enhancement associated with the unsteady airfoil motion magnifies with increasing pitch rate. Observing the flow over a range from $\alpha^+ = 0.05$ to 0.6, they determined that both the magnitude of maximum lift and the angle of attack at which it occurs (dynamic stall angle) increase with α^+ . As expected, an increase in the initial slope of the lift coefficient curve accompanies these trends. Francis and Keesee (1985) showed that the relationship between the maximum lift coefficient and α^+ over a pitch rate range from zero to 0.4 follows the same trend as the attack angle at which dynamic stall occurs. Strickland and Graham (1986) introduced a stall delay angle, defined as:

$$\Delta\alpha_{N \text{ stall}} = \alpha_{N \text{ dyn. stall}} - \alpha_{\text{static stall}}$$

where $\alpha_{N \text{ dyn. stall}}$ is the angle of attack corresponding to dynamic stall and $\alpha_{\text{static stall}}$ is the angle of attack corresponding to static stall. The occurrence of dynamic stall is based upon the occurrence of leading edge separation. They found that the stall delay angle at the nose, $\Delta\alpha_{N \text{ stall}}$, is proportional to the square root of the non-dimensional pitch rate. Based upon research that the maximum lift

coefficient follows the same trend as the pitch angle corresponding to dynamic stall, it can be deduced that $C_{L \text{ Max}}$ should follow the same trend. This has not been analyzed.

The connection between the surface velocity field over a constantly pitched airfoil and the non-dimensional pitch rate was investigated by Walker, Helin, and Strickland (1985) and by Walker and Chou (1987). Walker, Helin, and Strickland discovered reverse flow velocities directly under the dynamic stall vortex to be over $140\% U_{\infty}$ for an α^+ of 0.2 and over $210\% U_{\infty}$ for an α^+ of 0.6. It was also noted that the leading edge vortex initially produces a much higher reverse flow velocity than the trailing edge vortex does. Walker and Chou confirmed these results, and additionally disclosed that the upper surface velocity reaches a sub-peak corresponding to the point of dynamic stall vortex initiation, and then resumes an increasing trend rapidly to the primary velocity peak.

It has been demonstrated that the airfoil pivot location has a substantial effect on the unsteady flow field development. Helin and Walker (1985) found that as the pivot point moves downstream along the chord, the onset of dynamic stall is delayed. The subsequent rapid development and movement of the vortex over the airfoil has no appreciable qualitative variation with pivot location.

Stephen, et al (1989) confirmed that the flow develops independently of pitch axis location after the leading edge vortex forms. In their comparative study of a two-dimensional NACA 0015 airfoil pitched about axes from half a chord length forward of the airfoil to half a chord length aft of the airfoil, they determined that the size of the vortical disturbances are similar. This indicates that the vortex strength may be comparable in each case. Although moving the pitch axes aft does delay the flow development, the magnitude of the pressure peak and thus of the

aerodynamic pressure forces decreases. In both the non-dimensional pitch rates studied, $\alpha^+ = 0.1$ and 15 , the delay due to an increase in α^+ was greater than the delay due to changing pitch axis. The trade off between dynamic stall delay and enhanced airfoil performance has not been thoroughly investigated to date.

Although the two-dimensional mixing layer is in many aspects dissimilar from a two-dimensional unsteady airfoil flow, the fundamental growth mechanisms are analogous. As described at the beginning of this section, the first indication of disturbances in the viscous flow around a NACA 0015 airfoil pitched at a constant rate is when flow reversals appear near the surface at the rear of the airfoil. This reversal moves up the airfoil surface, and large eddies materialize in the interface (shear) layer between the inviscid free stream and the viscous boundary layer. It is these eddies, originated by the Kelvin-Helmholtz instabilities, which can be seen to combine and contribute to the growth of the unsteady separation region and the subsequent dynamic stall vortex growth (Albertson, 1989).

The plane mixing layer was first shown to contain large scale structures by the experiments of Brown and Roshko (1974). Winant and Browand (1974) added that the growth of the mixing layer was due to the propagation of the instability waves, which then roll up into discrete two-dimensional vortex structures. The subsequent growth of the mixing layer is due to the interaction between these large scale vortex structures, known as "pairing". Since this two-dimensional mixing layer is significantly simpler than bounded flows, a great deal of analysis has been accomplished to identify the means with which to control the growth and behavior of the large scale structures. The initial instability frequency has been shown to fluctuate in time by as much as 10%, and there is actually a broad band of frequencies present (Browand, 1986)

It has been conclusively demonstrated that the large scale vortex structures present in the mixing layer can be modified by external forcing. Ho and Huang (1982) and Oster and Wygnanski (1982) showed that the pairing interactions between the large scale structures (and thus mixing layer growth) can be modified by introducing coherent perturbations at the initiation of mixing. By controlling the pairing interactions, the spreading rate of the mixing layer can be enhanced or inhibited, in some extents even changing the signs of the Reynolds stress values and reducing turbulent energy. Ho and Huang specifically determined that the mechanisms by which vortex pairing was controlled depended upon the frequency at which the flow was controlled.

The methods of vortex pairing control have been extended to include the shear layer of a reattaching separated flow. Troutt, Scheelke, and Norman (1984) verified that large scale structures similar to those present in the mixing layer were also characteristic of the reattaching separated flow over a downstream facing step. Bhattacharjee, Scheelke, and Troutt (1986) and Roos and Kegelmann (1986) explored control of the flow over a downstream facing step using two separate control mechanisms. Bhattacharjee, Scheelke, and Troutt (1986) used a hot-wire probe to ascertain the characteristic frequency associated with the Kelvin-Helmholtz instabilities. Power spectra of the signal yielded a broad peak that gradually shifted towards lower frequencies with downstream position. This shift is accredited to the large-scale vortex amalgamations occurring in the separated shear layer. Forcing at the characteristic initial vortex passage produces a sharp spike in the power spectra at the natural flow frequency. This coincides with increased temporal and spatial correlation in the spanwise flow and a considerable reduction in reattachment length. Forcing between Strouhal numbers of 0.2 and 0.4 is the most effective

forcing range over a large range of Reynolds number. Strouhal number, St , is defined as the frequency, multiplied by the characteristic length and divided by the free stream velocity. Roos and Kegelmann (1986) confirmed these results using an oscillating flap at the step edge to excite the flow.

Several experimental analyses have been performed upon static airfoils under varying control conditions. Ahuja and Burrin (1984) used high frequency external acoustic control over a cambered airfoil to improve lift coefficients over 50% greater than natural conditions. Lift enhancement is highly dependent upon both frequency and amplitude, with the greatest lift increase occurring near 665 Hz. This frequency is directly proportional to the characteristic flow frequency.

By internally injecting acoustic forcing near the characteristic flow frequency, Collins (1981) was able to partially reattach the upper surface separation and increase lift coefficients by 20%. Maestrello (1986) confirmed this with external acoustic forcing, and in addition determined that velocity perturbation magnitudes in the region of transition could be reduced significantly.

Zaman, Bar-Sever, and Mangalam (1987) were able to remove laminar separation completely by low frequency ($St \leq 5$) external acoustic oscillations over a smooth airfoil. In addition, lift enhancements were achieved with large amplitude, high frequency excitation ($St = 4-25$) in the post-stall regime. Tunnel cross-resonances induce large transverse velocity fluctuations near the airfoil that enhance the separation control. However, these fluctuations would not be present in the open flow over an aircraft in actual flight. The authors suggest that the excitation mechanisms which produce reduced separation must hinge on the instability of the separated shear layer, but must also be influenced by the presence of the solid boundary and the separation location. Huang, Maestrello, and Bryant

(1987) discovered that the shear layer was extremely sensitive to sound excitation in the vicinity of the separation point. In their study of internal acoustic excitation of a static airfoil, they also verified that the most beneficial forcing frequencies are those of the instability waves. Forcing at the fundamental shedding frequency or its harmonic increases entrainment in the early part of the shear layer and drastically reduces the extent of separation.

The strong coupling between the injected sound and the shear layer instability frequency was confirmed by Huang, Maestrello, and Bryant (1987). They were able to reduce the region of separation over a symmetrical airfoil by injecting sound through the leading edge of the airfoil. At the same time, there is increased entrainment on the early part of the shear layer and increased circulation. The stall angle of attack is delayed and lift magnitude is enhanced significantly for extended attack angles.

Spanwise blowing has been successfully employed by numerous researchers to take advantage of the enhanced lift created in unsteady flows. The initial interest in this particular method of control arose from the need to manipulate the flow over static delta wings, which are characterized by continuing leading edge vortex growth and breakdown. Both Bradley and Wray (1974) and Campbell (1976) found that blowing a stream of high-pressure air over a wing surface, parallel to the leading edge, delayed leading edge vortex growth and thus the deleterious effects of vortex breakdown at the higher angles of attack. This delay is accompanied by lift increases of up to 50% over uncontrolled flows. The delay of vortex breakdown postpones static stall of the lifting surface as well. Seginer and Salomon (1986) used spanwise blowing over a canard-wing configuration for static angles of attack to augment both lift and lift-to-drag ratios and delay static stall.

Using a slightly different means, Roberts, et al (1985) and Wood and Roberts (1988) used tangential mass injection through a slot along the leading edge of a static delta wing. Direct control of the primary separation allows significant control of the vortex flow up to sixty degrees angle of attack. The primary effect of tangential leading edge blowing is to reduce the strength of the vortical flow, resulting in an extended regime of stable, controlled vortical flow over the upper surface of the wing. In addition, the separation line relocates to an inboard position under the forcing.

Successful attempts have been made to control unsteady flows such as the one discussed here. Carr and McAlister (1983) used a leading edge slat on an oscillating airfoil to produce a flow that remains attached to the airfoil for angles of attack well above those characteristic to the natural flow. The dynamic stall was significantly delayed, while at the same time the severity of the stall was reduced. Lutges, Robinson, and Kennedy (1985) were able to obtain similar results by introducing single air pulses through a two-dimensional slot located at $0.2c$ on an oscillating NACA 0015 surface. The flow control was effective only when the pulse corresponded to periods of high shear and large accumulations of vorticity.

The experiments reviewed here have revealed control techniques which are proven to reduce separation, delay static stall, and enhance lift over airfoils and wings. The current experimental analysis seeks to adapt the demonstrated methods of controlling shear flows to active forcing of dynamic stall flows. Active control of the flow over a constantly pitching airfoil should result in initial reduction in the unsteady separation region, dynamic stall delay, and extensions of the high-lift envelope.

4: DESCRIPTION OF RESEARCH PROCEDURES

The experiments discussed here were conducted in the Frank J. Seiler Research Laboratory open return, low speed wind tunnel at the U. S. Air Force Academy. The wind tunnel has a 3.0 ft x 3.0 ft test section. A NACA 0015 airfoil with a 6 in chord and a 2 ft span was pitched at constant rates from 0° to 50° about the 0.25c pitch axis. The four non-dimensional pitch rates, α^+ , of 0.01, 0.05, 0.1, and 0.15 were applied. The airfoil motion was accomplished using a stepper motor assembly controlled by a MassComp 5500 microcomputer system.

The flow was visualized by introducing smoke using a smoke wire. The smoke wire technique used by Helin and Walker (1985) was employed. Theatrical fog fluid was applied to a 0.005 in tungsten wire, leaving fine droplets almost uniformly along the wire. A current applied to the wire evaporated the oil, producing fine streaklines across the test section.

The visibility of the smoke depends upon the amount of light being scattered by the smoke particles themselves. Maximum efficiency results when the smoke is illuminated by direct light. If other parts of the test section are illuminated with too great a light level, the background reflections may overpower the smoke. The lighting for the wind tunnel was provided by synchronized strobe lights. The high intensity arc-lamp strobe lights are synchronized with a 35mm still camera and a 16mm high-speed camera to illuminate the streakline flow. The strobe lights have an average 7 μ s flash duration, which froze the flow.

The characteristic flow frequencies were obtained using two separate methods. To ascertain a preliminary frequency of the Kelvin-Helmholtz instability

waves, results from high-speed film at the four specified non-dimensional pitch rates was analyzed. The waves were visually counted with repeatable results. These frequencies were then verified by placing a hot-film probe near the surface of the flow and evaluating the signal through power spectra methods.

5: EXPERIMENTAL RESULTS

The first segment of the research project was to develop concise means of acoustic and tangential-pulsed air forcing applicable to a NACA 0015 airfoil under constant pitch motions. The tangential-pulsed air system is shown in Figure 2. Three spanwise, two dimensional slots 1/16 " wide and 18" long are placed in the airfoil at the leading edge, 20% chord, and 40% chord. The slots are configured so that the exiting air follows the airfoil surface for a period of time, creating control along the surface at the slot. The slots are supplied with air on each end. A timing circuit translates a signal from a square wave generator and simultaneously pulses six valves at a specified frequency between zero and sixty hertz. The control system allows either coincident pulsing through each slot or forcing from any given slot combination. A flow visualization study was performed to verify that the presence of the slots themselves did not significantly alter the flow.

The acoustic forcing system is shown in Figure 3. A 15", 220 watt woofer is placed in an enclosure so that the optimum sound intensity and quality exits through a 2.25" hole in the top of the casing. An additional sleeve padded with acoustic foam surrounds the speaker enclosure to dampen out resonant frequencies from the box itself. Sound is supplied to the test section through a series of rigid

pipes, which permit transmission of a low noise signal in a non-intrusive manner. The acoustic sound is focused on the leading portion of the airfoil, since forcing nearest to the origination of the instability waves yields the greatest effect. The speaker is controlled with a sine wave generator to produce frequencies as low as 15 hertz. The signal at frequencies lower than 15 hertz is inconsistent, and control required at these levels was provided by the pulsed air system. An attempt was made to acoustically force the flow internally through the slots designed for pulsed air blowing. Unfortunately, there was a problem with the transmission of the acoustic pressure wave, and sufficient acoustic amplitude did not reach the slots.

The next step in the experiment was to ascertain the characteristic frequencies associated with the Kelvin-Helmholtz instabilities. These frequencies were first visually measured from high speed films of the streak-line flow for the four different non-dimensional pitch rates. Although not a completely accurate method of frequency determination, this procedure allowed an initial estimate that aided in interpretation of the more accurate hot-film data. By placing a hot-film probe just outside the airfoil boundary layer, a clear trace of the instability induced waves was obtained. As expected, these traces resemble the form of a Doppler burst. Figure 4 shows examples of the natural signal, the signal acoustically forced at the characteristic frequency, and subsequently at the first subharmonic of the flow. The hot-film signals were first quantified using a storage oscilloscope, and subsequently using a power spectra analysis.

Examples of power spectra for the static case at an angle of attack of five degrees and the dynamic cases of $\alpha^+ = 0.01$ and 0.05 are given in Figure 5. The free stream velocity for these cases was 25 ft/sec. Three dominant frequencies are apparent in the static case at 20 Hz, 10 Hz, and 5 Hz. As the 20 Hz signal is the

most powerful, it indicates the natural frequency associated with the Kelvin-Helmholtz instability waves. The peaks at 5 Hz and 10 Hz show that pairing of the instability waves is occurring, resulting in the emergence of the subharmonics. The non-dimensional pitch rates of $\alpha^+ = 0.01$ and 0.05 yield strong peaks near 50 Hz, specifically at 52 Hz and 56 Hz, respectively.

Flow visualization at a non-dimensional pitch rate of 0.01 is shown in Figure 6 for the non-forced (natural) case and for the flow externally forced at the characteristic frequency of 15 Hz. Since the low freestream velocity of 10 ft/sec required for flow visualization did not allow accurate hot-film measurements, the characteristic instability frequencies were determined from high-speed film. It should be noted that the natural frequency of the flow decreased with freestream velocity. At attack angles lower than $\alpha = 20^\circ$, there are no discernible differences between the forced and natural flow. The two cases, columns one and two, differ significantly at this angle of attack. There is a loosely defined leading edge, or dynamic stall, vortex present on the upper airfoil surface for both the natural and forced flow. However, the vortex on the natural airfoil covers a larger portion of the upper airfoil surface than in the controlled case.

The differences between the flows have diminished by $\alpha = 25^\circ$, but it is still clear that the natural flow contains traces of large scale structures. Off the trailing edge, an upward trend in the flow coincides with a faint trailing edge vortex, characteristic of dynamic stall flows. There is also a less-defined vortex near the leading edge. The controlled flow shows no evidence of such structures. By $\alpha = 30^\circ$, the two flows again appear similar.

A flow visualization study was also conducted at a pitch rate of 0.05 using external acoustic forcing. Figure 7 shows both the natural case in column one and the forced case, 33 Hz, in column two. As at an α^+ of 0.01, the effects of the acoustic forcing are not evident prior to an angle of attack of 20° . At this attack angle, a slight disruption occurs in the forced alternating vortices shed off the trailing edge. The apparent accelerated mixing of the flow continues to an attack angle of 25° , where the acoustically forced case shows a separated region that is slightly smaller than the than the natural case. At $\alpha = 30^\circ$, a dynamic stall vortex is clearly developing on the natural airfoil, indicated by the region void of smoke near the leading edge. Although there is a similar void region near the leading edge of the forced airfoil, it is not as developed. In addition, the width of the upper surface separation region is reduced in the controlled case. By $\alpha = 35^\circ$ the two flows are similar.

It is forcing at the subharmonic frequency that has been most beneficial in controlling flows such as a wake or a mixing layer. In these flows, controlling the flow at the subharmonic frequency enhances vortex pairing and delays separation. It is reasonable to assume that forcing at the subharmonic frequency of an unsteady flow will yield similar results. Figure 8 shows the flow for several angles of attack under natural and external subharmonic control. The two flows are similar prior to $\alpha = 25^\circ$. At that attack angle, the controlled flow appears slightly more slightly more turbulent, especially near the trailing edge. In addition, there is no evidence of the depressed area of the separation region midway along the airfoil that indicates dynamic stall vortex formation in the natural case. At $\alpha = 30^\circ$, there is still no sign of a dynamic stall vortex in the forced flow, while one is clearly present on the natural airfoil. In addition, the overall width of the separation region near the

airfoil leading edge is reduced by approximately half in the controlled flow. By $\alpha = 35^\circ$, a dynamic stall vortex has formed on the forced airfoil as well and there are no discernible differences between the two cases.

6: CONCLUSIONS

The analysis of acoustically and tangentially controlled flow over a airfoil pitching at the constant non-dimensional rates of 0.01, 0.05, 0.1, and 0.15 has yielded the following conclusions:

1. Forcing the flow at the characteristic frequency delays dynamic stall vortex growth, and at the same time results in a decreased separation region over the upper airfoil surface.
2. Forcing the flow at the first subharmonic also delays dynamic stall vortex formation and growth, and in addition causes a greater reduction in the separation region over the surface than the case forced at the characteristic frequency.
3. The natural instability frequencies of the flow are not only dependent upon non-dimensional pitch rate, but also on free-stream velocity.

From these results, it can be expected that, by actively forcing the flow at

the first subharmonic or lower subharmonics, enhanced lift can be maintained for a longer period of time.

7: RECOMMENDATIONS

Based upon the results obtained in this study, the following recommendations for extended research are offered:

- 1) The study of the Kelvin-Helmholtz instability frequency should be expanded to include the effects of changing free stream velocity and actual pitch rate.
- 2) A thorough pressure representation of each case is imperative to assess the practical benefits of the active control.
- 3) The information gained from active control studies of these flows should be extended to include interactive forcing and airfoil motion controls, to allow a maximization of the benefits of active control.

REFERENCES

- Ahuja, K. K. and Burrin, R. H. (1984) "Control of Flow Separation by Sound," AIAA Paper No. 84-2298.
- Albertson, J. A., Troutt, T. R., Siuru, W. D., and Walker, J. M. (1987) "Dynamic Stall Vortex Development and the Surface Pressure Field of a Pitching Airfoil," AIAA Paper No. 87-1333.
- Albertson, J. A., Troutt, T. R., and Kedzie, C. R. (1988) "Unsteady Aerodynamic Forces at Low Airfoil Pitching Rates," AIAA Paper No. 88-2579.
- Bhattacharjee, S., Scheelke, B., and Troutt, T. R. (1986) "Modification of Vortex Interactions in a Reattaching Separated Flow," AIAA J., Vol. 24, No. 4, pp 623-629.
- Bradley, R. G. and Wray, W. O. (1974) "A Conceptual Study of Leading-Edge-Vortex Enhancement by Blowing," J. Aircraft, Vol. 11, No. 1, pp 33-38.
- Browand, F. K. (1986) "The Structure of the Turbulent Mixing Layer," Physica, Vol. 18D, pp 135-148.
- Brown, G. and Roshko, A. (1974) "On Density Effects and Large Structure in Turbulent Mixing Layers," J. Fluid Mech., Vol. 64, pp 693-704.
- Campbell, J. F. (1976) "Augmentation of Vortex Lift by Spanwise Blowing," J. Aircraft, Vol. 13, No. 9, pp 727-732.
- Carr, L. W. (1988) "Progress in Analysis and Production of Dynamic Stall," J. Aircraft, Vol 25, No. 1, pp 6-17.
- Carr, L. W. and McAlister, K. W. (1983) "The Effect of a Leading-Edge Slat on the Dynamic Stall of an Oscillating Airfoil," AIAA Paper No. 83-2533.
- Collins, F. G. (1981) "Boundary-Layer Control on Wings Using Sound and Leading-Edge Serration," AIAA J., Vol. 19, No. 2, pp 129-130.
- Cook, R. J. (1987) "Similarity Conditions for Flows about Pitching Airfoils," FJSRL-TM-87-001.
- Francis, M. S. and Keesee, J. E. (1985) "Airfoil Dynamic Stall Performance with Large-Amplitude Motions," AIAA J., Vol. 23, No. 11, pp 1653-1659.
- Gad-el-Hak, M. (1986) "The Use of the Bye-Layer Technique for Unsteady Flow Visualization," J. of Fluids Eng., Vol. 108, No. 3, pp 34-38.
- Gad-el-Hak, M. and Bushnell, D. (1991) "Status and Outlook of Flow Separation Control," AIAA Paper No. 91-0037.
- Ham, N. D. and Garelick, M. S. (1968) "Dynamic Stall Considerations in Helicopter Rotors," J. Am. Hel. Soc., Vol. 13, No. 2, pp 40-50.

Helin, H. E. and Walker, J. M. (1985) "Interrelated Effects of Pitch Rate and Pivot Point on Airfoil Dynamic Stall," AIAA Paper No. 85-0130.

Herbst, W. B. (1985-a) "Dynamics of Air Combat," J. Aircraft, Vol. 20, No. 7, pp 594-598.

Herbst, W. B. (1985-b) "Supermaneuverability," Workshop on Unsteady Separated Flows, pp 1-8.

Ho, C. M. and Huerre, P. (1984) "Perturbed Free Shear Layers," Ann. Rev. Fluid Mech., Vol. 16, pp 365-424.

Huang, L. S., Maestrello, L., and Bryant, T. D. (1987) "Separation Control Over an Airfoil at High Angles of Attack by Sound Emanating From the Surface," AIAA Paper No. 87-1261.

Jumper, E. J., Schreck, S. J., and Dimmick, R. L. (1987) "Lift-Curve Characteristics for an Airfoil Pitching at Constant Rate," J. Aircraft, Vol. 24, No. 10, pp 680-687.

Luttges, M. W., Robinson, M. C., and Kennedy, D. A. (1985) "Control of Unsteady Separated Flow Structures on Airfoils," AIAA Paper No. 85-0531.

Maestrello, L. (1986) "Active Transition Fixing and Control of the Boundary Layer in the," AIAA J., Vol. 24, No. 10, pp 1577-1581.

McCroskey, W. J. (1982) "Unsteady Airfoils," Annual Review of Fluid Mechanics, Vol. 25, No. 1.

Nelson, C. F., Koga, D. J., and Eaton, J. K. (1989) "Control of the Unsteady, Separated Flow Behind an Oscillating, Two-Dimensional Flap," AIAA Paper No. 89-1027.

Oster, D. and Wignanski, I. (1982) "The Forced Mixing Layer Between Parallel Streams," J. Fluid Mech., Vol. 123, pp 91-130.

Roberts, L., Hesselink, L., Kroo, I., and Wood, N. J. (1985) "The Control of Vortical Flows Over a Delta Wing," Workshop - II on Unsteady Separated Flows, FJSRL-TR- pp 317-322.

Roos, F. W. and Kegelman, J. T. (1986) "Influence of Excitation on Coherent Structures in Reattaching Turbulent Shear Layers," AIAA Paper No. 86-0112.

Seginer, A. and Salomon, M. (1986) "Performance Augmentation of a 60-Degree Delta Aircraft Configuration by Spanwise Blowing," J. Aircraft, Vol. 23, No. 11, pp 801-807.

Stephen, E., et al (1989) "Extended Pitch Axis Effects on the Flow about Pitching Airfoils," AIAA Paper 89-0025

Strickland, J. H. and Graham, G. M. (1986) "Dynamic Stall Inception Correlation for Airfoils Undergoing Constant Pitch Rate Motions," AIAA J., Vol. 24, No. 4, pp 678-680.

Troutt, T. R., Scheelke, B., and Norman, T. R. (1984) "Organized Structures in a Reattaching Separated Flow Field," J. Fluid Mech., Vol. 143, pp 413-427.

Walker, J. M. and Chou, D. C. (1987) "Forced Unsteady Vortex Flows Driven by Pitching Airfoils," AIAA Paper No. 87-1331.

Walker, J. M., Helin, H. E., and Chou, D. C. (1985) "Unsteady Surface Pressure Measurements on a Pitching Airfoil," AIAA Paper No. 85-0532.

Walker, J. M., Helin, H. E., and Strickland, J. H. (1985) "An Experimental Investigation of an Airfoil Undergoing Large Amplitude Pitching Motions," AIAA Paper No. 85-0039.

Whitford, R. (1987) Design For Air Combat, Jane's Publishing Company Limited, New York, NY.

Winant, C. D. and Browand, F. K. (1974) "Vortex Pairing: The Mechanism of Turbulent Mixing-Layer Growth at Moderate Reynolds Number," J. Fluid Mech., Vol. 63, Part 2, pp 237-255.

Wood, N. J. and Roberts, L. (1988) "Control of Vortical Lift on Delta Wings by Tangential Leading-Edge Blowing," J. Aircraft, Vol. 25, No. 3, pp 236-243.

Zaman, K. M. B. Q., Bar-Sever, A., and Mangalam, S. M. (1987) "Effect of Acoustic Excitation on the Flow over a Low-Re Airfoil," J. Fluid Mech., Vol. 182, pp 127-148.

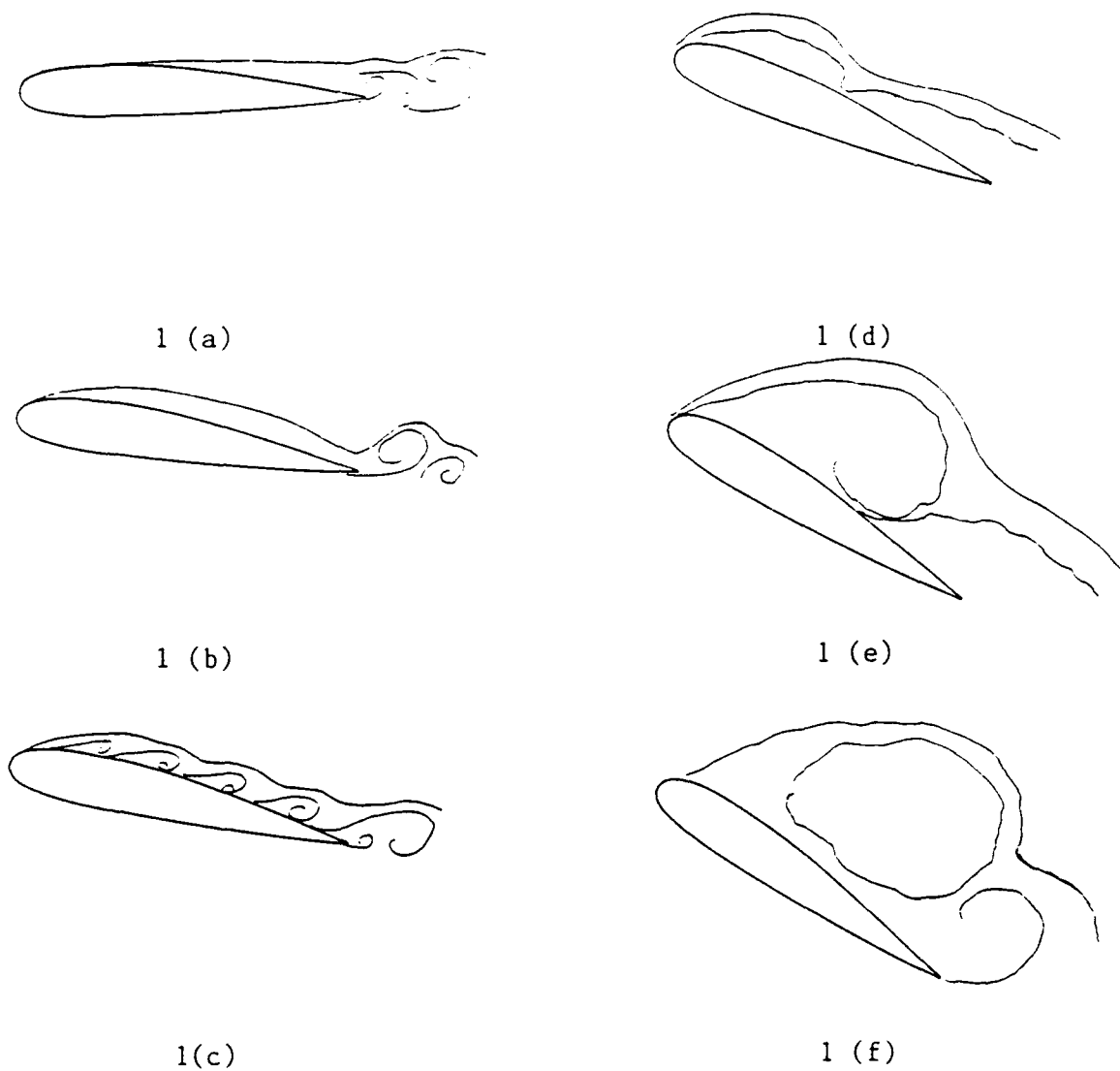


FIGURE 1. Evolution of the unsteady boundary layer and development of the dynamic stall vortex.

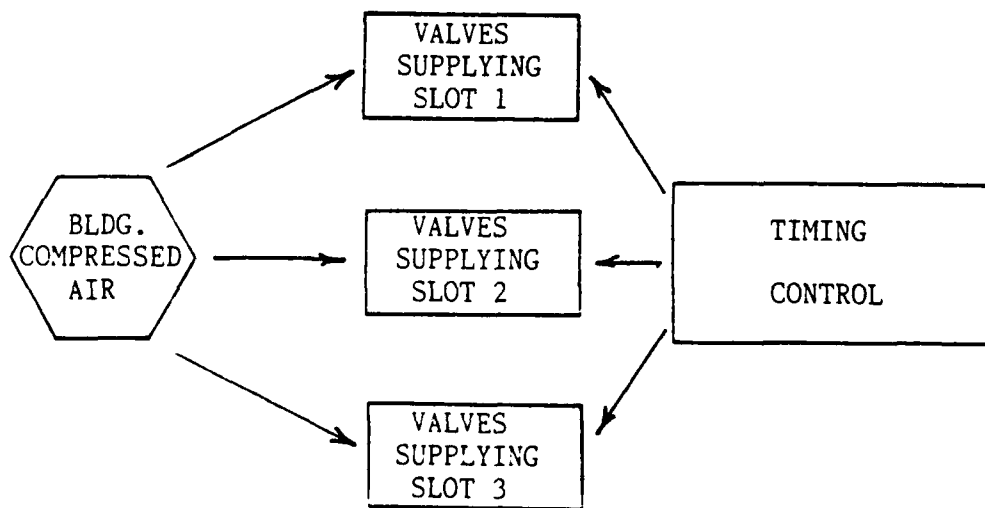
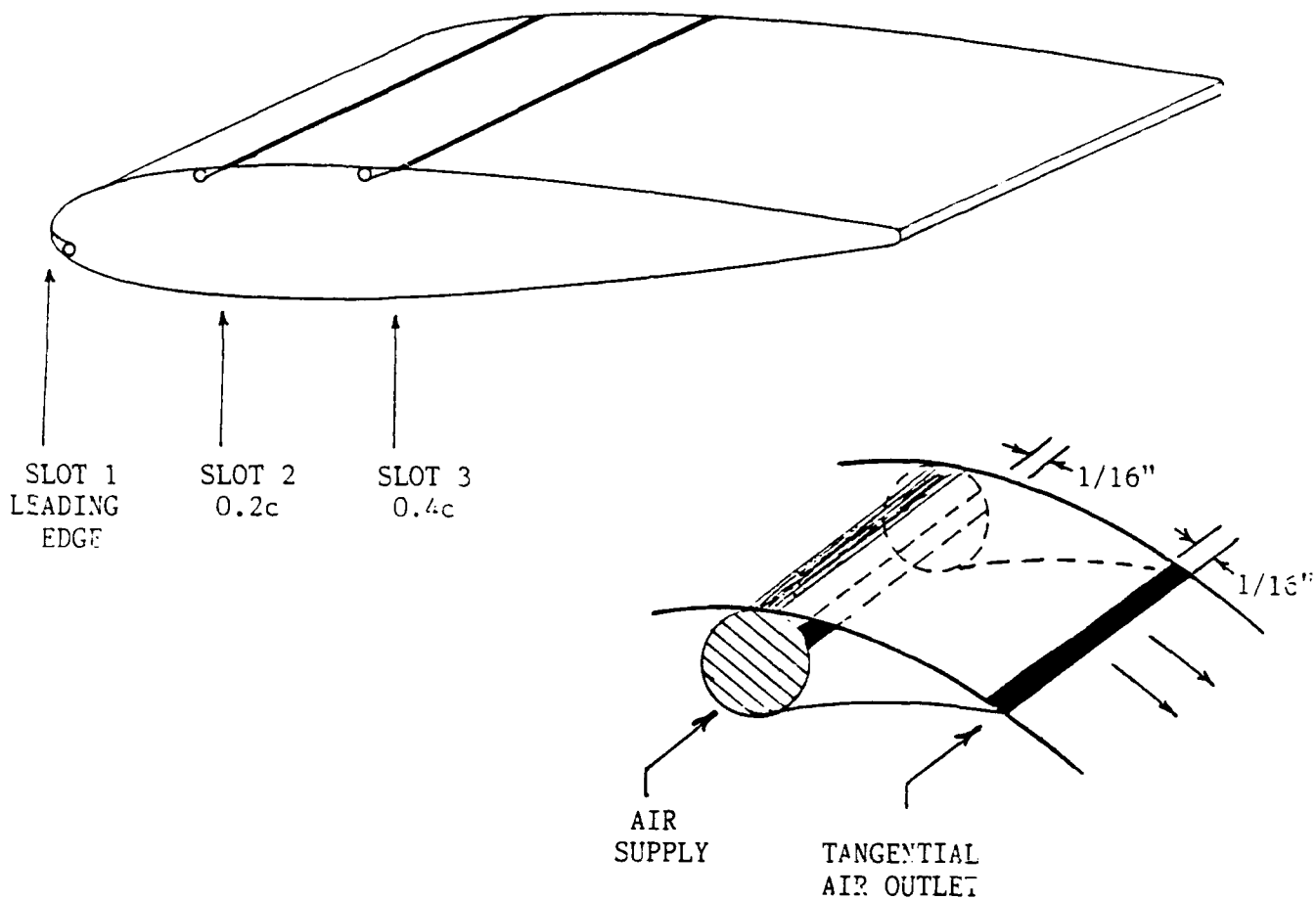


FIGURE 2. Tangential Pulsed Air Blowing System

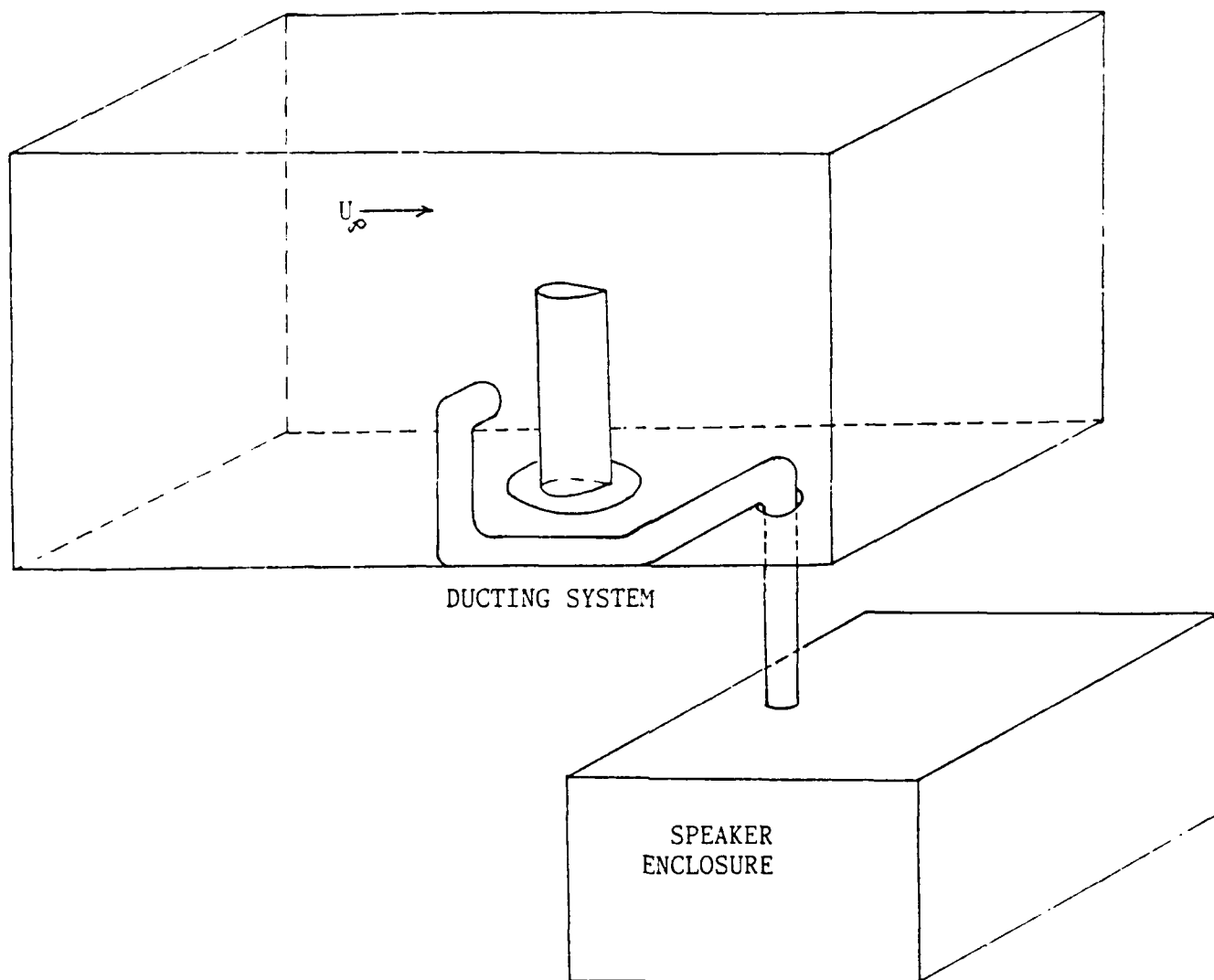
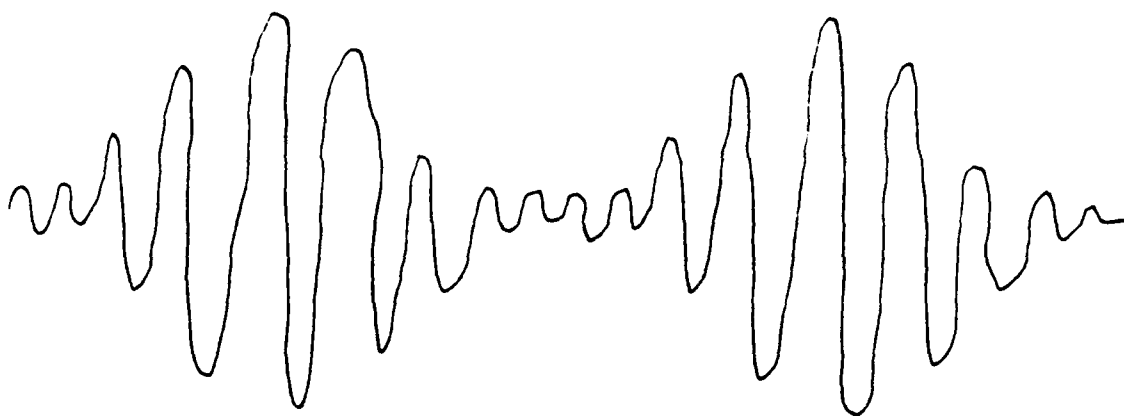
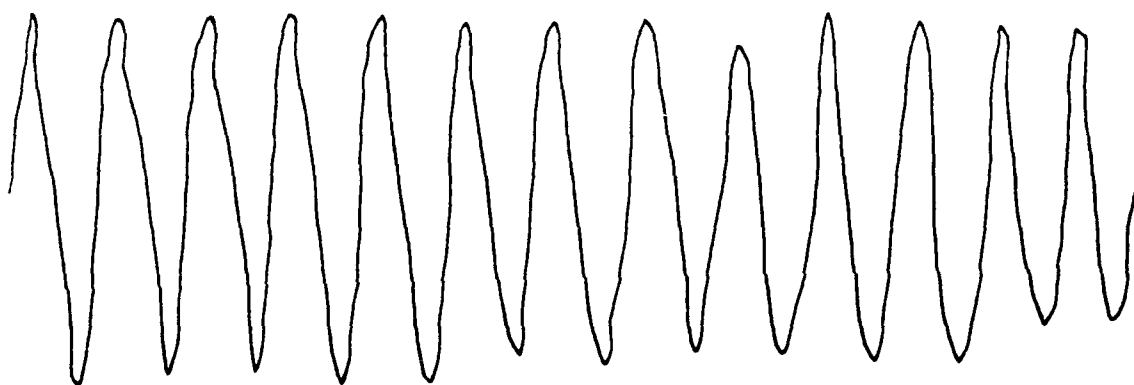


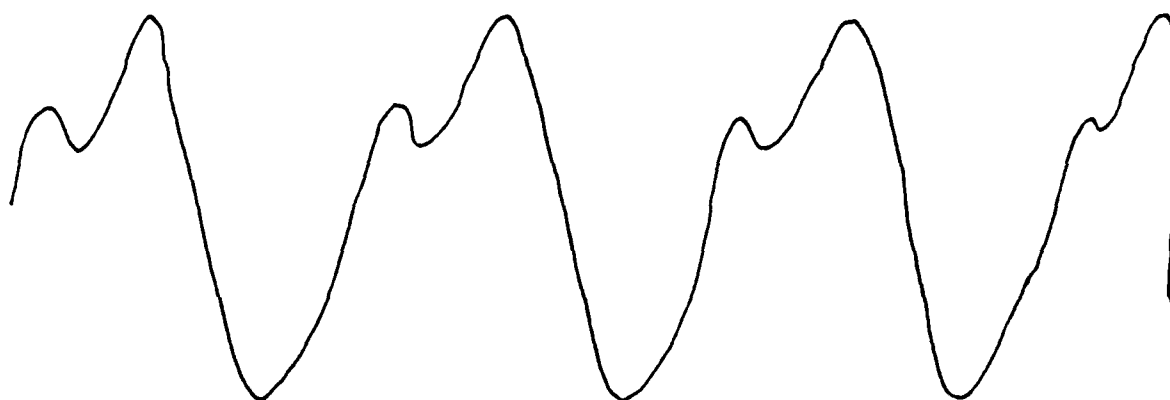
FIGURE 3. External Acoustic Control System



4(a): Natural Signal

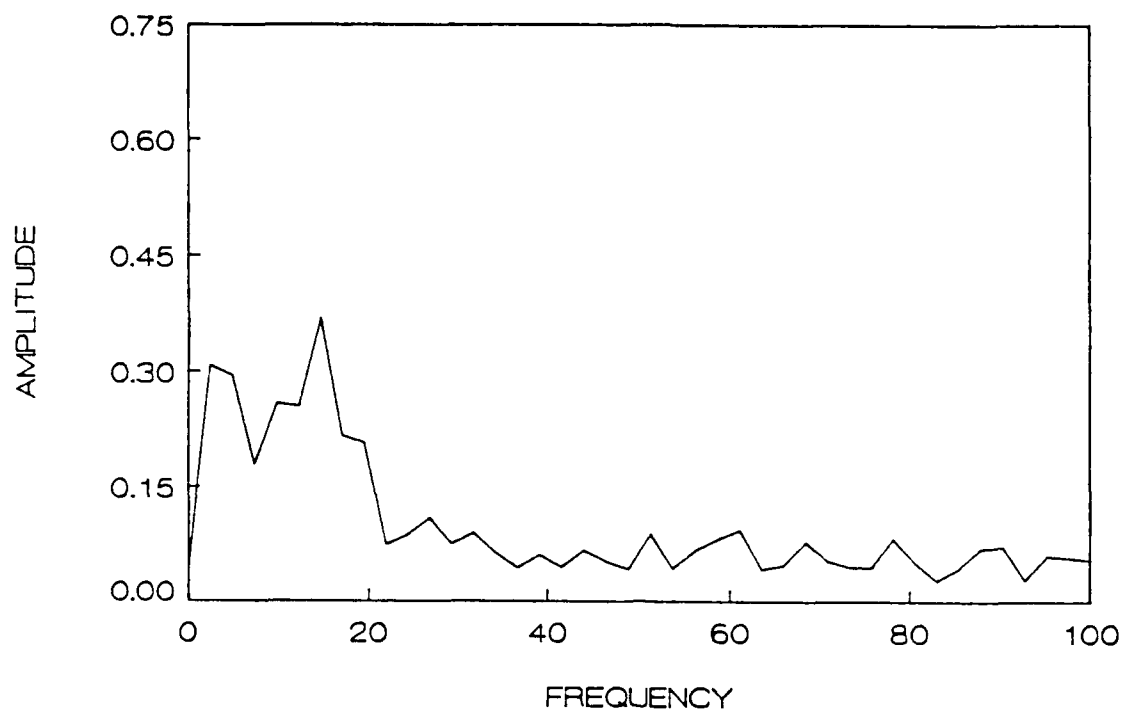


4(b): Signal forced at natural frequency

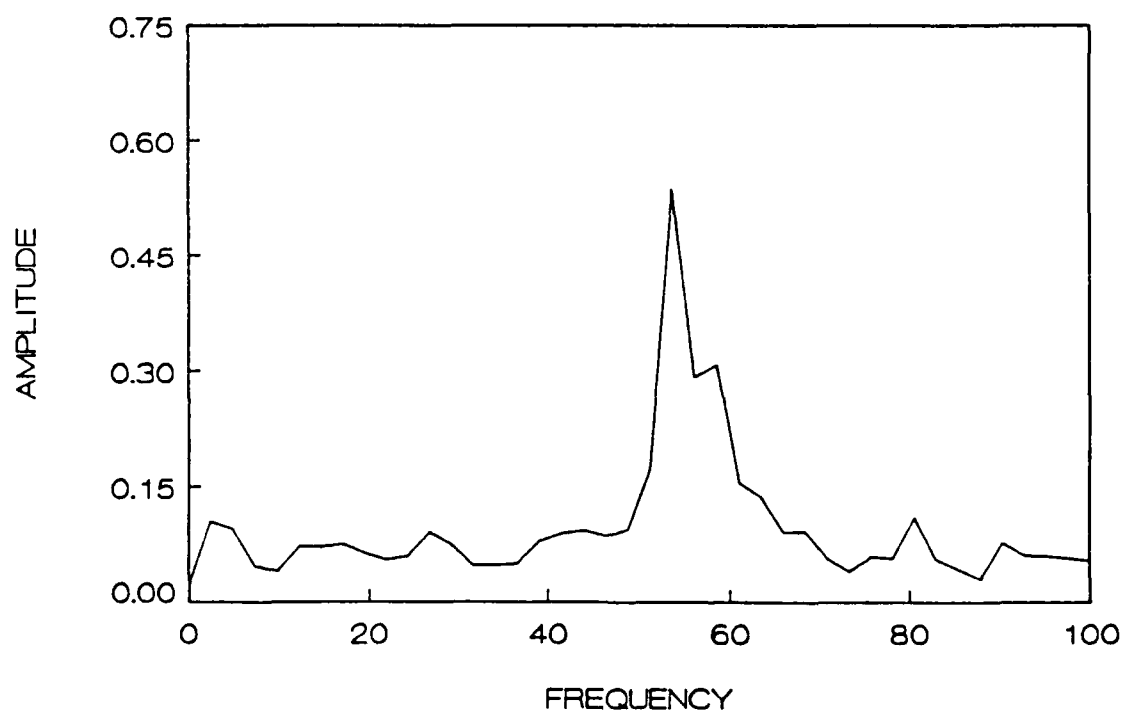


4(c): Signal forced at subharmonic frequency

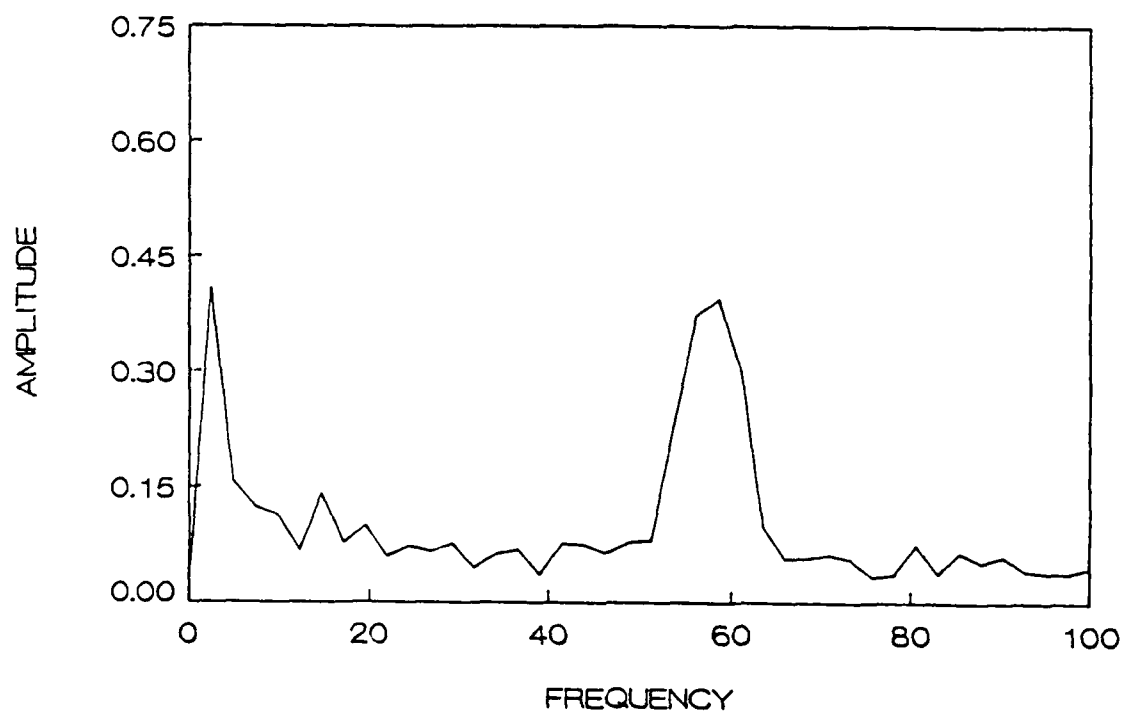
FIGURE 4. Hot-Film Signals from an unsteady airfoil flow



5(a): Static Power Spectra, ≈ 5 degrees.



5(b): Dynamic Power Spectra, $\dot{\theta} = 0.01$



5(c): Dynamic Power Spectra, $\tau = 0.05$

FIGURE 5. Power Spectra for hot-film data, static and dynamic airfoil flow.



0 HZ

10°

15°

20°

25°

30°

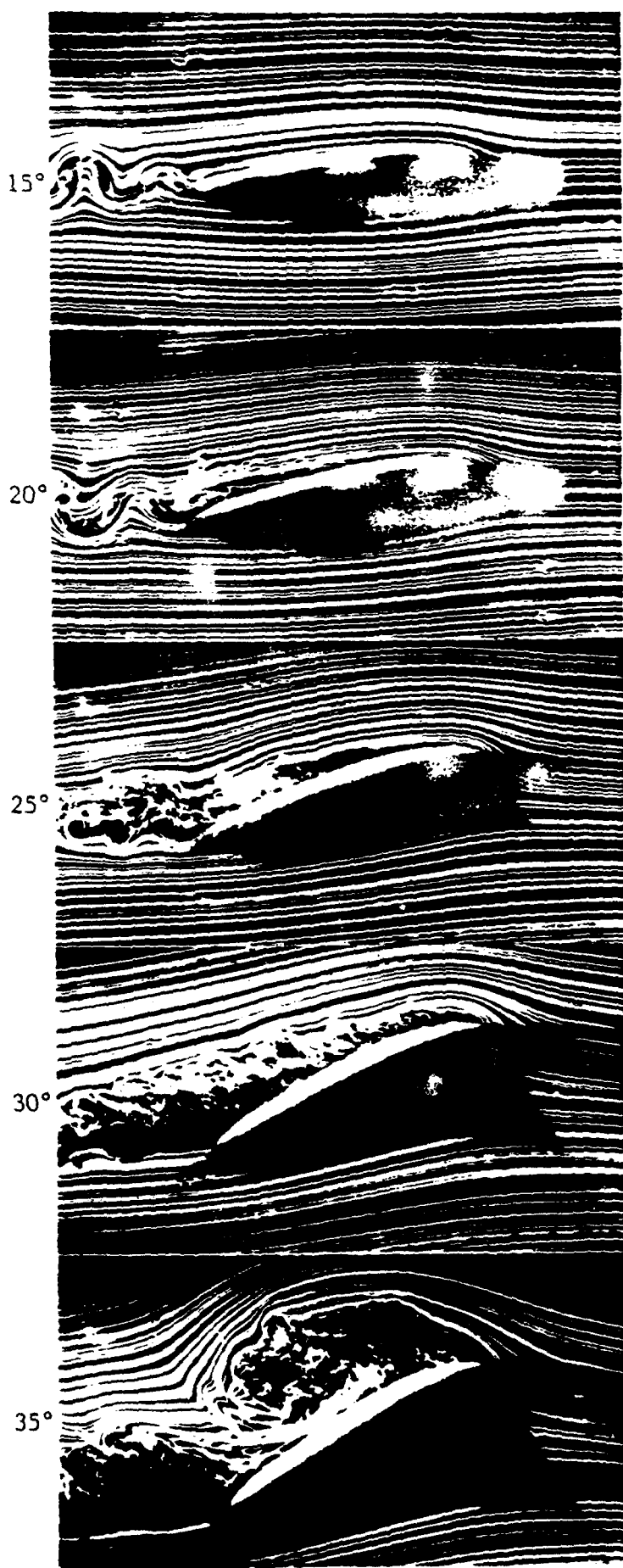


15 HZ

Figure 6. Flow visualization at $\alpha^+ = 36.35$ for natural flow (0 Hz) and flow forced at the characteristic frequency (15 Hz).



0 HZ



15°

20°

25°

30°

35°

15 HZ

FIGURE 7. Comparison of Visualized Flow Under Natural Conditions and Under Subharmonic Forcing. $\epsilon = 0.05$.



0 HZ

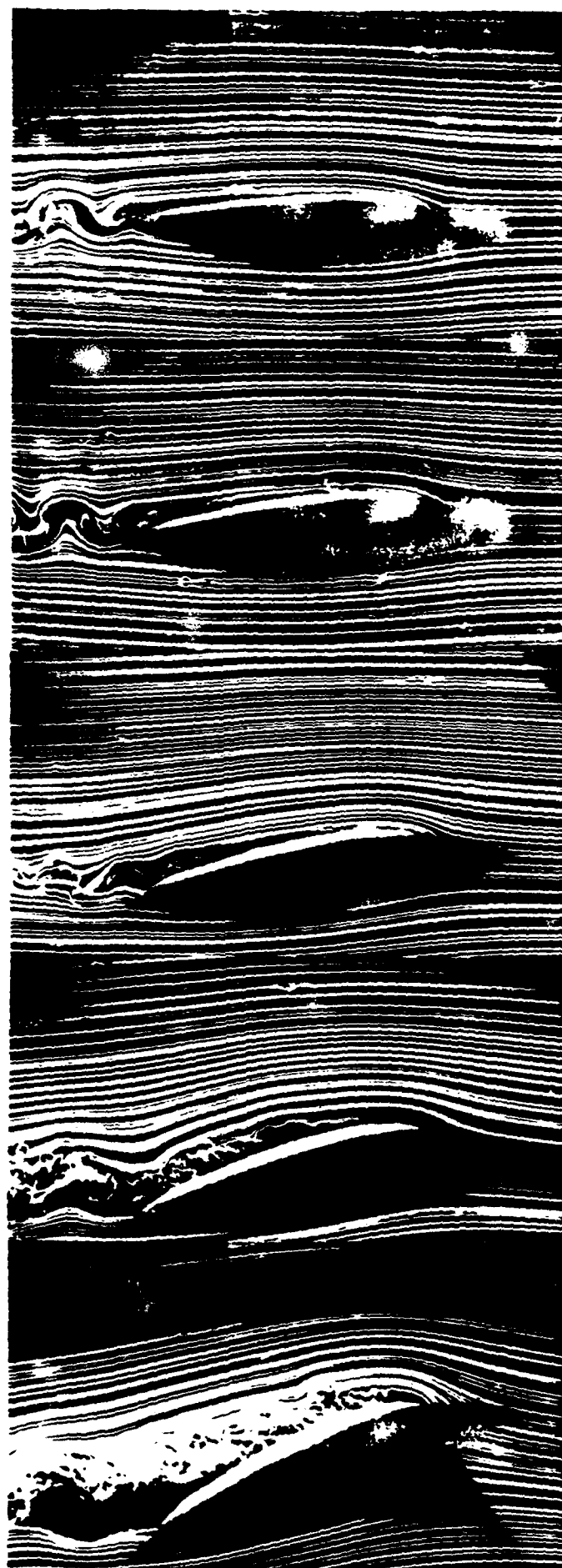
10°

15°

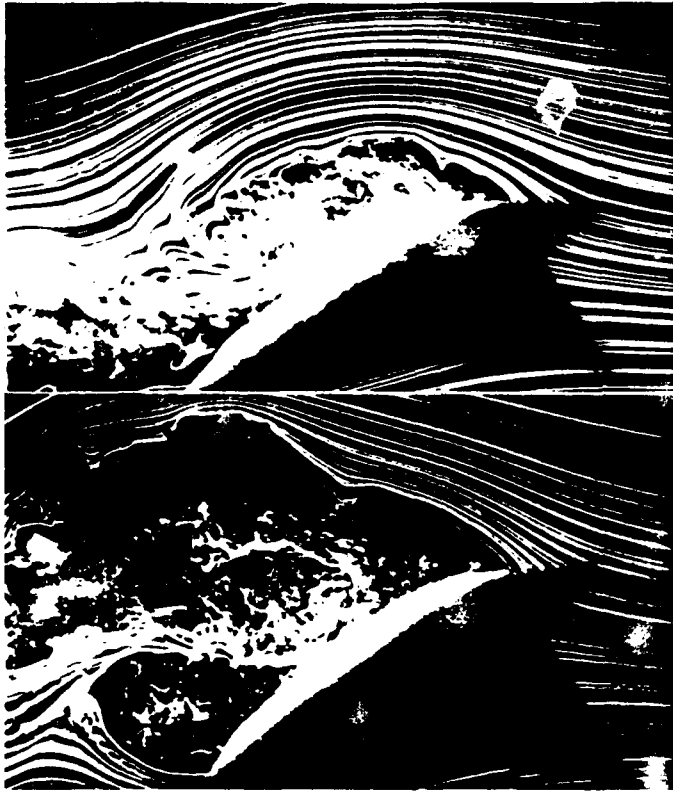
20°

25°

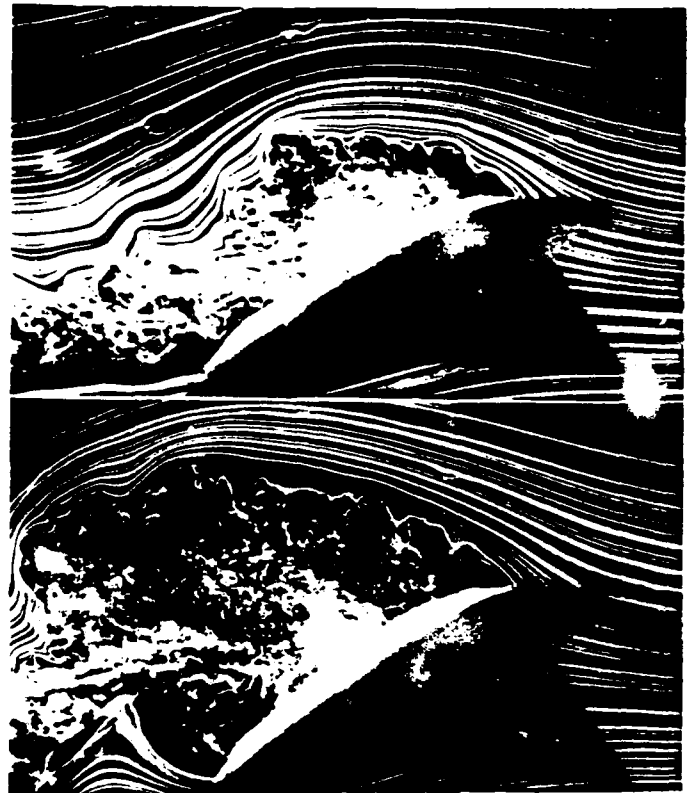
30°



33 HZ



0 HZ



33 HZ

Figure 8. Flow visualization at $\alpha^+ = 0.05$ for natural flow (0 Hz) and flow forced at the characteristic frequency (33 Hz).

1990 USAF-UES RESEARCH INITIATION PROGRAM (RIP)

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

UNIVERSAL ENERGY SYSTEMS, INC.

FINAL REPORT

MODELING AND CONTROL OF A FUNDAMENTAL STRUCTURE-
CONTROL SYSTEM: A CANTILEVER BEAM AND A STRUCTURE-
BORNE REACTION-MASS ACTUATOR

Prepared by: Hung V. Vu, Ph.D., Principal Investigator
Department of Mechanical Engineering
California State University, Long Beach
Long Beach, CA 90840

Date: December 20, 1990

Contract No: F49620-88-C-0053/SB5881-0378

Purchase Order No: S-210-10MG-021

MODELING AND CONTROL OF A FUNDAMENTAL STRUCTURE-
CONTROL SYSTEM: A CANTILEVER BEAM AND A STRUCTURE-
BORNE REACTION-MASS ACTUATOR

Hung V. Vu *

California State University, Long Beach

Gwocheang O. Shaw **

University of California, Irvine

ABSTRACT

Many problems of large space structures (LSS) in structural dynamics and control can be addressed by studying fundamental beam systems. In this investigation, a system consisting of a cantilever beam and a structure-borne reaction-mass actuator (RMA) is studied. In the first part, the work is focused on the modeling of the system in which exact Euler-Bernoulli beam equation is used to provide accurate model. The RMA is considered as a part of the boundary condition. The natural frequencies, mode shapes, and forced response are determined. In the second part, both linear optimal regulator and full-order estimator are designed. The complete control system is synthesized by connecting the optimal linear quadratic regulator (LQR) and the full-order estimator.

* Assistant Professor, Department of Mechanical Engineering

** Graduate Student, Department of Mechanical Engineering

ACKNOWLEDGEMENTS

I wish to thank the Air Force Office of Scientific Research, Bolling AFB, DC, and the Frank J. Seiler Research Laboratory, USAF Academy, Colorado for sponsorship of this research. I would also like to thank the Universal Energy Systems, Inc. for all administrative aspects of this research program.

CHAPTER 1

Introduction

The investigation is broken into two parts. First, the modeling and analysis are carried out to study the system dynamics. The system considered is a cantilever beam with a structure-borne reaction-mass actuator (RMA) attached at the free end. The applied forcing function is arbitrary. The governing partial differential equation of motion of the Euler-Bernoulli beam in transverse vibration is derived. The RMA is modeled as two-point concentrated masses connected by a spring and a viscous damper. The undamped natural frequencies and the corresponding normalized mode shapes are determined by solving the eigenvalue problem where the RMA (undamped) is considered as a part of the boundary condition. The forced responses for both the undamped and damped cases are obtained by means of modal analysis. Both frequency response and transient response are considered. The system under consideration is classified as passive vibration control whose vibration suppression capability is limited compared to active vibration control.

In the second part, a feedback control system is designed to suppress the vibration in an active vibration control fashion. The control system design is based on modern control theory where state feedback, output feedback, pole placement, performance index, etc. are considered. Important aspects of control system design such as stability, performance, controllability, observability, reconstructability, etc. are studied.

The control design is broken into three independent steps. The first step is a regulator design in which the complete state is assumed available for feedback. A performance

index is defined for the design goal. By minimizing the performance index, the optimal linear quadratic Gaussian regulator (LQR) can be found. Usually the assumption of complete-state feedback is not practical, an estimator has to be used to reconstruct the state in this case. This is the second step. The last step is to connect the regulator and estimator to complete the design (LQG design). The complete system is called output feedback control system.

Chapter 2 is devoted to free vibration problem. A frequency equation is obtained by solving the eigenvalue problem. Because it is a transcendental equation, a numerical algorithm is employed to calculate the natural frequencies which are the roots of the frequency equation. Each natural frequency is related to a vibration mode which has a particular mode shape. The orthogonality derived shows that each mode shape is orthogonal to the others.

The forced response is calculated in Chapter 3. Lagrange's equation, which takes the approach of energy consideration, is used to derive the differential equation of motion. By making use of the assumption of modal analysis and the orthogonality, a set of uncoupled ordinary differential equations are obtained. The solution of the forced response is in analytic form. Modal analysis approach is also applied to the damped system to have a set of coupled equations. However, an analytic solution is not available to these equations. If we put these equations in state-space form, a software package can offer the numerical solution. The state-space equation is also convenient for control analysis. But modal analysis transforms the continuous system into an infinite-degree-of-freedom system in terms of the generalized coordinates, which actually can not be measured for

feedback. This problem can be solved by using a transformation matrix which converts the equation back to the physical coordinate system. The physical coordinates are chosen at positions along the beam where the vibration needs to be suppressed. The state-space equation is nondimensionalized for numerical calculations. Several common parameters are defined. Case study is at the end of Chapter 3.

Chapter 4 is focused on the regulator design. The state-space differential equation derived is the basic model for control theory analysis. The dynamics of the reaction mass actuator (RMA) is discussed. The stability and controllability also have been analyzed to make sure the design is possible. In the fourth section of this chapter, optimal feedback gains are calculated. The sensitivity analysis shows that the control system has enough ability to compensate the parameter changes in the system modeling.

In Chapter 5, a full-order estimator is studied. The purpose of introducing the estimator into the control loop is to reconstruct the state of the system when the complete state feedback is not available. Also, the estimator offers certain degree of filtering of the noise which always exists in the practical implementation. The reconstructability guarantees that the estimator design is possible. The estimator feedback gains are calculated. The output feedback control system is set up by combining the regulator and estimator. The characteristics of the complete system totally depends on the regulator and estimator. Therefore, satisfactory response should be expected if regulator and estimator are well designed.

Case study is based on a simple model. First two modes are included in this model which requires the state feedback from two positions along the beam. We choose $x = 0.5L$ and $x = L$ for sensors' positions. The numerical results are presented in the case study. A FORTRAN code is written and used to solve for the natural frequencies and calculate the coefficients of the differential equations. A software package, MATRIX_X, is used to calculate the system response and analyze the control system. The FORTRAN and MATRIX_X programs are listed in the Appendix.

CHAPTER 2

Modeling of Cantilever Beam-RMA System

Frequency Equation and Natural Frequencies

The problem of determining the natural frequencies of a system is called the eigenvalue problem. For transverse vibrations of a beam (Fig. 2.1), Euler-Bernoulli's model is used to obtain the governing partial differential equation. In the eigenvalue problem, the damping and external forces, $f(x, t)$ and $F(t)$, are set to zero.

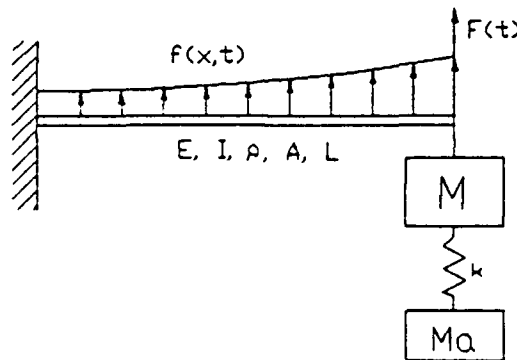


Figure 2.1 Cantilever Beam-RMA System

The Euler-Bernoulli's beam equation is derived by considering an infinitesimal element dx of a beam and the internal shear forces and moments acting on it. The coordinates and free-body diagram are shown in Fig. 2.2. The beam theory provides us with the following relations:

$$M_b = EI(x) \frac{\partial^2 w(x, t)}{\partial x^2} \quad (2-1a)$$

$$V(x, t) = -\frac{\partial}{\partial x} \left[EI(x) \frac{\partial^2 w(x, t)}{\partial x^2} \right] \quad (2-1b)$$

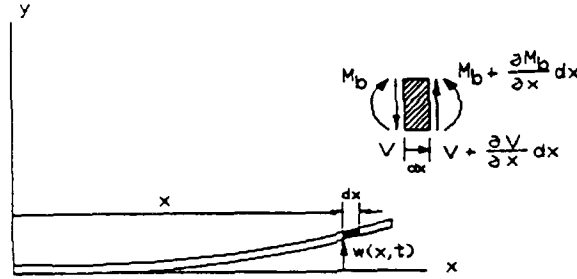


Figure 2.2 Free-Body Diagram of Element dx

The mass of the beam element is $\rho A dx$ and the acceleration of the element is $\partial^2 w / \partial t^2$. Applying Newton's second law, we have

$$\sum F_y = \rho A dx \frac{\partial^2 w}{\partial t^2} \quad (2-2)$$

$$-V + V + \frac{\partial V}{\partial x} dx = \rho A \frac{\partial^2 w}{\partial t^2} dx \quad (2-3)$$

Combining Eqs. 2-1 and 2-3, we obtain

$$\frac{\partial^2}{\partial x^2} \left[EI(x) \frac{\partial^2 w(x, t)}{\partial x^2} \right] + \rho A(x) \frac{\partial^2 w(x, t)}{\partial t^2} = 0 \quad (2-4)$$

If the beam has uniform cross section, Eq. 2-4 reduces to

$$\frac{\partial^4 w(x, t)}{\partial x^4} + \frac{\rho A}{EI} \frac{\partial^2 w(x, t)}{\partial t^2} = 0 \quad (2-5)$$

Euler-Bernoulli's beam equation, Eq. 2-5, describes the transverse vibrations of a continuous beam.

The boundary conditions at $x = 0$ are easily obtained as

$$w(0, t) = 0 \quad (2-6)$$

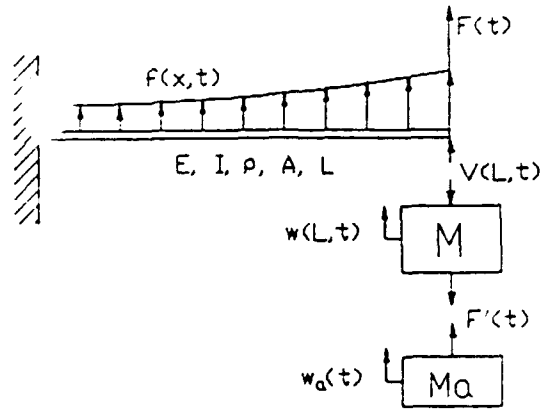


Figure 2.3 Free-Body Diagram at $x=L$

$$\frac{\partial w(0, t)}{\partial x} = 0 \quad (2-7)$$

The RMA in this system can be modelled as two point-masses connected by a spring. The free-body diagram at $x = L$ is shown in Fig. 2.3. Because the moment at $x = L$ is zero, by using Eq. 2-1 we have

$$EI \frac{\partial^2 w(L, t)}{\partial x^2} = 0 \quad (2-8)$$

Applying Newton's law to both masses, M and M_a , we obtain

$$M \frac{\partial^2 w(L, t)}{\partial t^2} = -V(L, t) - F'(t) \quad (2-9)$$

$$M_a \frac{\partial^2 w_a(t)}{\partial t^2} = F'(t) \quad (2-10)$$

The shear force $V(L, t)$ is equal to $-EI \partial^3 w(L, t) / \partial x^3$ in Eq. 2-1 and the spring force $F'(t)$ is equal to $k[w(L, t) - w_a(t)]$. Substituting these two relations into Eqs. 2-9 and 2-10, we get

$$EI \frac{\partial^3 w(L, t)}{\partial x^3} = M \frac{\partial^2 w(L, t)}{\partial t^2} + k[w(L, t) - w_a(t)] \quad (2-11)$$

$$M_a \frac{\partial^2 w_a(t)}{\partial t^2} = k[w(L, t) - w_a(t)] \quad (2-12)$$

Eqs. 2-6, 2-7, 2-8, 2-11, and 2-12 are the five boundary conditions of the system which can be used with the governing equation, Eq. 2-5 to solve the eigenvalue problem.

The modal analysis approach assumes that the solution of Eq. 2-5 is expressed as

$$w(x, t) = \phi(x)e^{j\omega t} \quad (2-13a)$$

$$w_a(t) = \phi_a e^{j\omega t} \quad (2-13b)$$

where $\phi(x)$ and ϕ_a , together, are the mode shape (also called mode function) of the system which will be determined in the third section of this chapter.

Differentiating Eq. 2-13a with respect to x and t , we obtain the following derivatives:

$$\frac{\partial^4 w(x, t)}{\partial x^4} = \frac{d^4 \phi(x)}{dx^4} e^{j\omega t} = \phi''''(x) e^{j\omega t} \quad (2-14)$$

$$\frac{\partial^2 w(x, t)}{\partial t^2} = -\phi(x) \omega^2 e^{j\omega t} \quad (2-15)$$

Introducing Eqs. 2-14 and 2-15 into Eq. 2-5, we have

$$\phi''''(x) - \frac{\rho A}{EI} \omega^2 \phi(x) = 0 \quad (2-16)$$

Let us define a constant λ as follows

$$\lambda^4 = \frac{\rho A}{EI} \omega^2 \quad (2-17)$$

Later in this section, we will see that λ is actually the eigenvalue of the system.

Using Eq. 2-17, Eq. 2-16 can be written as

$$\phi''''(x) - \lambda^4 \phi(x) = 0 \quad (2-18)$$

By using modal analysis approach, we have simplified the partial differential equation, Eq. 2-5, into a fourth-order ordinary differential equation, Eq. 2-18, which has the general solution:

$$\phi(x) = A_1 \cosh \lambda x + A_2 \cos \lambda x + A_3 \sinh \lambda x + A_4 \sin \lambda x \quad (2-19)$$

Also, substituting Eq. 2-13 into the boundary conditions, Eqs. 2-6, 2-7, 2-8, 2-11, and 2-12 and canceling the term $e^{j\omega t}$, we obtain

$$\phi(0) = 0 \quad (2-20)$$

$$\phi'(0) = 0 \quad (2-21)$$

$$\phi''(L) = 0 \quad (2-22)$$

$$EI\phi'''(L) = -M\omega^2 \phi(L) + k[\phi(L) - \phi_a] \quad (2-23)$$

$$-M_a\omega^2 \phi_a = k[\phi(L) - \phi_a] \quad (2-24)$$

The important relation between ϕ_a and $\phi(L)$ can be derived directly from Eq. 2-24, which is

$$\phi_a = \frac{k}{k - M_a\omega^2} \phi(L) = \frac{\omega_a^2}{\omega_a^2 - \omega^2} \phi(L) \quad (2-25)$$

where $\omega_a^2 = k/M_a$. Substituting Eq. 2-25 into 2-23, we get

$$\phi'''(L) = -\frac{1}{EI} \left[M\omega^2 + M_a \frac{\omega^2 \omega_a^2}{\omega_a^2 - \omega^2} \right] \phi(L) \quad (2-26)$$

The coefficient term on the right-hand side of Eq. 2-26 can be arranged into dimensionless form. Let us introduce the following dimensionless parameters:

$$\alpha = \frac{M}{\rho AL} \quad (\text{mass ratio of } M \text{ to the beam mass}) \quad (2-27a)$$

$$\mu = \frac{M_a}{M} \quad (\text{mass ratio of the point-masses}) \quad (2-27b)$$

$$f_a = \frac{\omega_a}{\sqrt{\frac{EI}{\rho AL^4}}} \quad (\text{tuning ratio of the RMA}) \quad (2-27c)$$

Eq. 2-26 is arranged into the following form

$$\phi'''(L) = -L \frac{\rho A \omega^2}{EI} \left[\frac{M}{\rho AL} + \frac{M}{\rho AL} \frac{M_a}{M} \frac{\frac{\omega^2}{EI}}{\frac{\omega_a^2}{EI} - \frac{\omega^2}{EI}} \right] \phi(L) \quad (2-28)$$

By substituting the dimensionless parameters, Eq. 2-27, into Eqs. 2-28 and 2-25, we have

$$\phi'''(L) = -\frac{(\lambda L)^4}{L^3} \alpha \left[1 + \mu \frac{f_a^2}{f_a^2 - (\lambda L)^4} \right] \phi(L) \quad (2-29)$$

$$o_a = \left[\frac{f_a^2}{f_a^2 - (\lambda L)^4} \right] \phi(L) \quad (2-30)$$

The four coefficients, A_1, A_2, A_3, A_4 in Eq. 2-19 are to be solved by using Eqs. 2-20, 2-21, 2-22, and 2-29.

Differentiating Eq. 2-19 three times with respect to x , we have

$$\phi(x) = A_1 \cosh \lambda x + A_2 \cos \lambda x + A_3 \sinh \lambda x + A_4 \sin \lambda x \quad (2-19)$$

$$\phi'(x) = \lambda(A_1 \sinh \lambda x - A_2 \sin \lambda x + A_3 \cosh \lambda x + A_4 \cos \lambda x) \quad (2-31)$$

$$\phi''(x) = \lambda^2(A_1 \cosh \lambda x - A_2 \cos \lambda x + A_3 \sinh \lambda x - A_4 \sin \lambda x) \quad (2-32)$$

$$\phi'''(x) = \lambda^3(A_1 \sinh \lambda x + A_2 \sin \lambda x + A_3 \cosh \lambda x - A_4 \cos \lambda x) \quad (2-33)$$

Applying Eqs. 2-20 and 2-21 to Eq. 2-19 and 2-31, respectively, we get

$$A_1 + A_2 = 0, \quad A_2 = -A_1 \quad (2-34)$$

$$A_3 + A_4 = 0, \quad A_4 = -A_3 \quad (2-35)$$

Applying Eqs. 2-22 and 2-29 to Eq. 2-32 and 2-33, respectively, we have

$$\lambda^2(A_1 \cosh \lambda L - A_2 \cos \lambda L + A_3 \sinh \lambda L - A_4 \sin \lambda L) = 0 \quad (2-36)$$

$$\begin{aligned} \lambda^3(A_1 \sinh \lambda L + A_2 \sin \lambda L + A_3 \cosh \lambda L - A_4 \cos \lambda L) = \\ -\frac{(\lambda L)^4}{L^3} \alpha \left[1 + \mu \frac{f_a^2}{f_a^2 - (\lambda L)^4} \right] \\ (A_1 \cosh \lambda L + A_2 \cos \lambda L + A_3 \sinh \lambda L + A_4 \sin \lambda L) \end{aligned} \quad (2-37)$$

Let us define

$$\begin{aligned} C &= \cosh \lambda L & c &= \cos \lambda L \\ S &= \sinh \lambda L & s &= \sin \lambda L \\ \mathcal{A} &= \alpha(\lambda L) \left[1 + \mu \frac{f_a^2}{f_a^2 - (\lambda L)^4} \right] \end{aligned}$$

Introducing these symbols into Eqs. 2-36 and 2-37 with Eqs. 2-34 and 2-35, we obtain

$$(C + c)A_1 + (S + s)A_3 = 0 \quad (2-38)$$

$$(S - s)A_1 + (C + c)A_3 = -\mathcal{A}[(C - c)A_1 + (S - s)A_3] \quad (2-39)$$

Rearranging Eqs. 2-38 and 2-39, a set of equations will be obtained to be solved for A_1 and A_3 .

$$\begin{bmatrix} (C + c) & (S + s) \\ (S - s) + \mathcal{A}(C - c) & (C + c) + \mathcal{A}(S - s) \end{bmatrix} \begin{bmatrix} A_1 \\ A_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (2-40)$$

In order to get non-trivial solutions, the determinant of the coefficient matrix in Eq. 2-40 has to be equal to zero, i. e.,

$$\begin{vmatrix} (C + c) & (S + s) \\ (S - s) + \mathcal{A}(C - c) & (C + c) + \mathcal{A}(S - s) \end{vmatrix} = 0 \quad (2-41)$$

Expanding the determinant we get

$$(C^2 + 2Cc + c^2) + \mathcal{A}(CS - Cs + cS - cs) - (S^2 - s^2) + \mathcal{A}(CS + Cs - cS - cs) = 0$$

By using the relations, $C^2 - S^2 = 1$ and $c^2 + s^2 = 1$, the above equation can be written as

$$Cc + 1 - \mathcal{A}(sC - cS) = 0 \quad (2-42)$$

Substituting the original terms into Eq. 2-40 and dividing Eq. 2-24 by Cc , we obtain

$$1 + \frac{1}{\cosh \lambda L \cos \lambda L} - \alpha(\lambda L) \left[1 + \mu \frac{f_a^2}{f_a^2 - (\lambda L)^4} \right] (\tan \lambda L - \tanh \lambda L) = 0 \quad (2-43)$$

The above equation is the frequency equation of the system which can be solved for the eigenvalue, λ . There is an infinite number of roots satisfying this equation and each $\lambda_n (n = 1, 2, \dots, \infty)$ is related to a vibration mode of the system where the corresponding natural frequency is

$$\omega_n = (\lambda_n L)^2 \sqrt{\frac{EI}{\rho AL^4}} \quad n = 1, \dots, \infty \quad (2-44)$$

If $f_a = 0$, i.e., the stiffness of the RMA is reduced to zero, then the system (Fig. 2.3) becomes a system shown in Fig. 2.4. The frequency equation for this system can be derived easily from Eq. 2-43 by substituting $f_a = 0$.

$$1 + \frac{1}{\cosh \lambda L \cos \lambda L} - \alpha(\lambda L)(\tan \lambda L - \tanh \lambda L) = 0 \quad (2-45)$$

or

$$1 + \frac{1}{\cosh \lambda L \cos \lambda L} - \frac{M\omega^2}{EI\lambda^3}(\tan \lambda L - \tanh \lambda L) = 0 \quad (2-46)$$

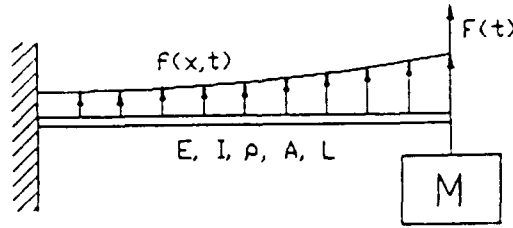


Figure 2.4 Cantilever Beam with Point-Mass M only

Solving the Frequency Equation

From the result of the preceding section, we know that the natural frequencies of the system are obtained by solving the frequency equation, Eq. 2-43.

$$1 + \frac{1}{\cosh \lambda L \cos \lambda L} - \alpha(\lambda L) \left[1 + \mu \frac{f_a^2}{f_a^2 - (\lambda L)^4} \right] (\tan \lambda L - \tanh \lambda L) = 0 \quad (2-43)$$

Eq. 2-43 is a transcendental equation. It is impossible to solve this equation analytically. A numerical method has to be employed to solve for the roots. An algorithm called Modified Linear Interpolation Method [1] is used for solving this equation.

Before writing the FORTRAN computer code, we have to arrange the equation to avoid the arithmetic overflow problem. Multiplying Eq. 2-43 by

$$\cos \lambda L \left[\frac{f_a^2 - (\lambda L)^4}{(\lambda L)^4 + 1} \right]$$

we obtain

$$\left[\frac{f_a^2 - (\lambda L)^4}{(\lambda L)^4 + 1} \right] \left(\cos \lambda L + \frac{1}{\cosh \lambda L} \right) - \alpha(\lambda L) \left[\frac{(1 + \mu)f_a^2 - (\lambda L)^4}{(\lambda L)^4 + 1} \right] (\sin \lambda L - \cos \lambda L \tanh \lambda L) = 0 \quad (2-47)$$

The program is listed in the Appendix. The transcendental equation 2-47 is programmed in the FUNCTION subroutine. For a different equation, only this subroutine needs to be modified. Also, the program has loops for different cases of tuning in the RMA which are set in the program constants FMIN, DF, and FMAX. The output of the program are the eigenvalues multiplied by the length of the beam. The number of eigenvalues depends on the frequency scanning range which is set by the constants XMIN, DX, and XMAX. For each eigenvalue, the corresponding natural frequency is calculated by Eg. 2-44.

If we set the following parameters as

$$\begin{aligned}\alpha &= 0.25, \quad \mu = 1.0 \\ XMIN &= 0.001, \quad DX = 0.1, \quad XMAX = 15. \\ f_a &= 0, \quad 1.8, \quad 17.6\end{aligned}$$

the eigenvalues calculated are

case	f_a	$\lambda_1 L$	$\lambda_2 L$	$\lambda_3 L$	$\lambda_4 L$	$\lambda_5 L$	$\lambda_6 L$
1	0.0	1.57	4.23	7.28	10.4	13.5	—
2	1.8	1.23	1.72	4.23	7.28	10.4	13.5
3	17.6	1.42	3.99	4.91	7.30	10.4	13.5

The first case is $f_a = 0$ which means that there is no mass M_a . This special case is the system shown in Fig. 2-4. In the second and third cases the RMA is tuned to the first and second modes of the first case. We can see in the second case that the first mode of case 1 is split into two modes with little effect on the others. The third case tuning to the second mode of case 1 shows the same phenomenon.

Orthogonality and Normalized Mode Shapes

From the first section of this chapter, we see that there are infinite vibration modes related to a continuous system. In this section we are going to determine

the mode shape for each vibration mode. Also, we are going to show that mode shapes satisfy the orthogonality. By using the orthogonality, the mode shapes are normalized.

Based on the assumption of modal analysis approach, the solution of a free transverse vibration of the system is

$$w(x, t) = \phi(x)e^{j\omega t} \quad (2-13a)$$

$$w_a(t) = \phi_a e^{j\omega t} \quad (2-13b)$$

where $\phi(x)$ and ϕ_a are the mode shapes and have been obtained in the first section as

$$\phi(x) = A_1 \cosh \lambda x + A_2 \cos \lambda x + A_3 \sinh \lambda x + A_4 \sin \lambda x \quad (2-19)$$

$$\phi_a = \left[\frac{f_a^2}{f_a^2 - (\lambda L)^4} \right] \phi(L) \quad (2-30)$$

Also, from the first section, we can have

$$A_2 = -A_1 \quad (2-34)$$

$$A_4 = -A_3 \quad (2-35)$$

And, Eq. 2-38 can be rewritten as

$$\frac{A_3}{A_1} = -\frac{\cosh \lambda L + \cos \lambda L}{\sinh \lambda L + \sin \lambda L} \quad (2-48)$$

Substituting these relations in Eq. 2-19 and 2-30, and denoting the n^{th} mode with subscript n , we can have the mode shape,

$$\phi_n(x) = C_n h_n(x) \quad (2-48a)$$

$$(\phi_a)_n = a_n \phi_n(L) \quad (2-48b)$$

where

$$h_n(x) = (\cosh \lambda_n x - \cos \lambda_n x) - \left[\frac{\cosh \lambda_n L + \cos \lambda_n L}{\sinh \lambda_n L + \sin \lambda_n L} \right] (\sinh \lambda_n x - \sin \lambda_n x)$$

$$a_n = \frac{f_a^2}{f_a^2 - (\lambda_n L)^4}$$

and C_n are arbitrary constants to be determined later.

Next, we will show that each mode shape in Eq. 2-48 satisfies the orthogonality condition. We start to derive the orthogonality in a general way (the cross section of the beam may not be uniform). Now, we go back to the differential equation of the system which is

$$\frac{\partial^2}{\partial x^2} \left[EI(x) \frac{\partial^2 w(x, t)}{\partial x^2} \right] + \rho A(x) \frac{\partial^2 w(x, t)}{\partial t^2} = 0 \quad (2-4)$$

Substituting Eq. 2-13a into 2-4, we obtain

$$\frac{d^2}{dx^2} \left[EI(x) \frac{d^2 \phi(x)}{dx^2} \right] = \rho A(x) \omega^2 \phi(x) \quad (2-49)$$

The m^{th} and n^{th} mode shapes must satisfy Eq. 2-49. Hence,

$$\frac{d^2}{dx^2} \left[EI(x) \frac{d^2 \phi_m(x)}{dx^2} \right] = \rho A(x) \omega_m^2 \phi_m(x) \quad (2-50)$$

$$\frac{d^2}{dx^2} \left[EI(x) \frac{d^2 \phi_n(x)}{dx^2} \right] = \rho A(x) \omega_n^2 \phi_n(x) \quad (2-51)$$

Multiplying Eq. 2-50 by $\phi_n(x)$, and integrating with respect to x from 0 to L , we obtain

$$\int_0^L \phi_n(x) \frac{d^2}{dx^2} \left[EI(x) \frac{d^2 \phi_m(x)}{dx^2} \right] dx = \int_0^L \rho A(x) \omega_m^2 \phi_m(x) \phi_n(x) dx \quad (2-52)$$

Integrating the left-hand side of Eq. 2-52 by parts and using the boundary conditions Eqs. 2-20, 2-21, 2-22, we get

$$\begin{aligned} \phi_n(x) \frac{d}{dx} \left[EI(x) \frac{d^2 \phi_m(x)}{dx^2} \right] \Big|_0^L - \int_0^L \frac{d\phi_n(x)}{dx} \frac{d}{dx} \left[EI(x) \frac{d^2 \phi_m(x)}{dx^2} \right] dx \\ = \omega_m^2 \int_0^L \rho A(x) \phi_m(x) \phi_n(x) dx \end{aligned} \quad (2-53a)$$

$$\begin{aligned} \phi_n(L) EI(x) \frac{d^3 \phi_m(L)}{dx^3} - \frac{d\phi_n(x)}{dx} EI(x) \frac{d^2 \phi_m(x)}{dx^2} \Big|_0^L \\ + \int_0^L EI(x) \frac{d^2 \phi_n(x)}{dx^2} \frac{d^2 \phi_m(x)}{dx^2} dx = \omega_m^2 \int_0^L \rho A(x) \phi_m(x) \phi_n(x) dx \end{aligned} \quad (2-53b)$$

$$\begin{aligned} \phi_n(L) EI(x) \frac{d^3 \phi_m(L)}{dx^3} + \int_0^L EI(x) \frac{d^2 \phi_n(x)}{dx^2} \frac{d^2 \phi_m(x)}{dx^2} dx \\ = \omega_m^2 \int_0^L \rho A(x) \phi_m(x) \phi_n(x) dx \end{aligned} \quad (2-53c)$$

Starting from Eq. 2-51 and repeating a similar derivation, we can have the following result:

$$\begin{aligned} \phi_m(L) EI \frac{d^3 \phi_n(L)}{dx^3} + \int_0^L EI \frac{d^2 \phi_n(x)}{dx^2} \frac{d^2 \phi_m(x)}{dx^2} dx \\ = \omega_n^2 \int_0^L \rho A \phi_m(x) \phi_n(x) dx \end{aligned} \quad (1-54)$$

Subtracting Eq. 2-54 from 2-53c, we obtain

$$0 = (\omega_m^2 - \omega_n^2) \int_0^L \rho A(x) \phi_m(x) \phi_n(x) dx + EI(x) [\phi_m(L) \phi_n'''(L) - \phi_n(L) \phi_m'''(L)] \quad (2-55)$$

The relation between $\phi(L)$ and $\phi'''(L)$ is given by

$$\phi'''(L) = -\frac{1}{EI(x)} \left[M\omega^2 + M_a \frac{\omega^2 \omega_a^2}{\omega_a^2 - \omega^2} \right] \phi(L) \quad (2-26)$$

With this relation, the second term on the right-hand side of Eq. 2-55 can be arranged as follows

$$\begin{aligned} & EI(x) [\phi_m(L)\phi_n'''(L) - \phi_n(L)\phi_m'''(L)] \\ &= M(\omega_m^2 - \omega_n^2) + M_a(\omega_m^2 - \omega_n^2) \left[\frac{\omega_a^2}{(\omega_m^2 - \omega_n^2)} \left(\frac{\omega_m^2}{\omega_a^2 - \omega_m^2} - \frac{\omega_n^2}{\omega_a^2 - \omega_n^2} \right) \right] \\ &= (\omega_m^2 - \omega_n^2) \left[M + M_a \frac{\omega_a^4}{(\omega_a^2 - \omega_m^2)(\omega_a^2 - \omega_n^2)} \right] \phi_m(L)\phi_n(L) \end{aligned} \quad (2-56)$$

Substituting this result back to Eq. 2-55, we come to the final equation as

$$\begin{aligned} 0 = (\omega_m^2 - \omega_n^2) \left\{ \int_0^L \rho A(x) \phi_m(x) \phi_n(x) dx + \right. \\ \left. \left[M + \frac{M_a \omega_a^4}{(\omega_a^2 - \omega_m^2)(\omega_a^2 - \omega_n^2)} \right] \phi_m(L) \phi_n(L) \right\} \end{aligned} \quad (2-57)$$

From Eq. 2-57, we can get the orthogonality condition of the system as

$$\begin{aligned} & \int_0^L \rho A(x) \phi_m(x) \phi_n(x) dx + \\ & \left[M + M_a \frac{\omega_a^4}{(\omega_a^2 - \omega_m^2)(\omega_a^2 - \omega_n^2)} \right] \phi_m(L) \phi_n(L) = 0, \quad (m \neq n) \end{aligned} \quad (2-58a)$$

$$\int_0^L \rho A(x) \phi_n^2(x) dx + \left[M + M_a \frac{\omega_a^4}{(\omega_a^2 - \omega_n^2)^2} \right] \phi_n^2(L) \neq 0, \quad (m = n) \quad (2-58)$$

When $m = n$, Eq. 2-58 is equal to a positive quantity. We can normalize the mode shapes by taking this quantity as unity. If we consider the case that the cross section of the beam is uniform, by taking the term ρAL out of the integral in

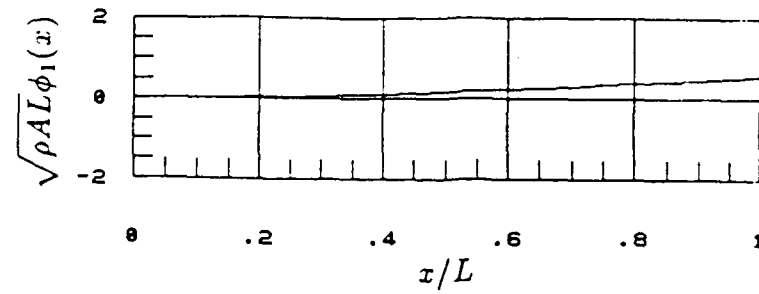
Eq. 2-58 and arranging the term in the bracket with Eq. 2-27 into dimensionless parameters, we have

$$\rho AL \left[\int_0^L \phi_m(x) \phi_n(x) (dx/L) + \alpha(1 + \mu a_m a_n) \phi_m(L) \phi_n(L) \right] = \delta_{mn} \quad (2-59)$$

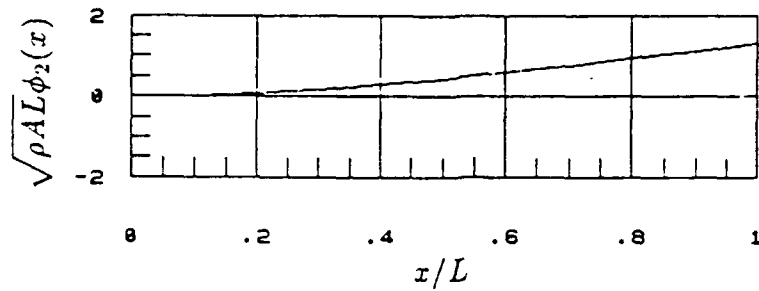
where δ_{mn} is the Kronecker delta. The mode shapes satisfying Eq. 2-59 are referred to as the normalized mode shapes. The coefficients C_n (see Eq. 2-48a) of the normalized mode shapes can be derived from Eq. 2-59 as

$$C_n = \left\{ \rho AL \left[\int_0^L h_n^2(x) (dx/L) + \alpha(1 + \mu a_n^2) h_n^2(L) \right] \right\}^{-1/2} \quad (2-60)$$

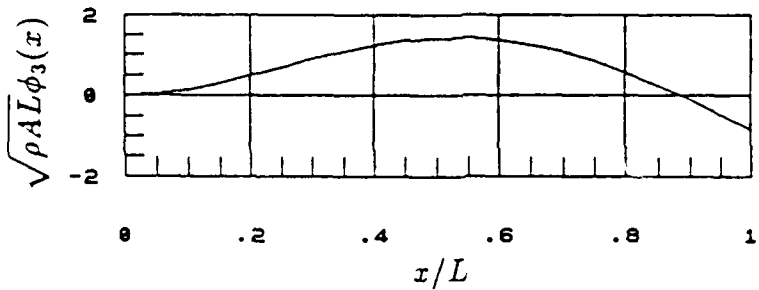
The first four normalized mode shapes for the case of $f_a = 1.8$ are shown in Fig. 2.5.



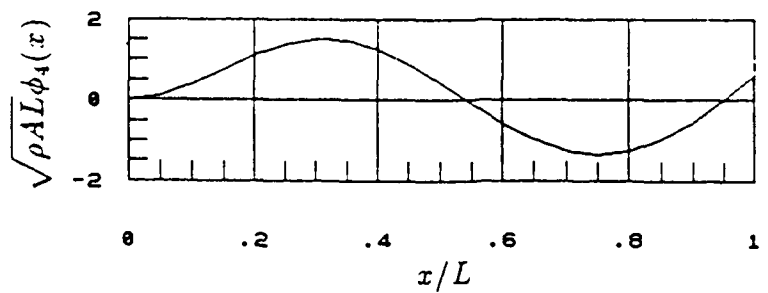
(a)



(b)



(c)



(d)

Figure 2.5 Normalized Mode Shapes (a) First Normalized Mode Shape
 (b) Second Normalized Mode Shape (c) Third Normalized Mode Shape
 (d) Fourth Normalized Mode Shape

CHAPTER 3

Forced Response of the System

Forced Response of the Undamped System

In this section we study the response of the undamped system subjected to the sinusoidal input. Referring to Fig. 2.1, if the external forces applied to the system are not equal to zero, the system response subjected to these forces is called the forced response. The solution of the forced response is assumed in the form of eigenfunction expansion,

$$\begin{aligned}w(x, t) &= \sum_{n=1}^{\infty} \phi_n(x) q_n(t) \\w_a(t) &= \sum_{n=1}^{\infty} a_n \phi_n(L) q_n(t)\end{aligned}\tag{3-1}$$

where the eigenfunctions, $\phi_n(x)$ and $a_n \phi_n(L)$, are the same as the mode shapes which have been solved in Chapter 2 and the generalized coordinates, $q_n(t)$, will be determined in this section.

We take the approach of energy consideration to derive the differential equation for the generalized coordinates $q_n(t)$. Also, the orthogonality derived in Section 2.3 will be used to simplify the derivation. Starting from Lagrange's equation expressed in terms of the generalized coordinates, we can write the differential equations as follows

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_n} \right) - \frac{\partial T}{\partial q_n} + \frac{\partial V}{\partial q_n} = Q_n(t) \quad n = 1, \dots, \infty\tag{3-2}$$

where T and V are the total kinetic and potential energy of the system respectively, and $Q_n(t)$ are the generalized nonconservative forces.

For the system considered, the kinetic energy of the beam is

$$T_b(t) = \frac{1}{2} \int_0^L \rho A \left[\frac{\partial w(x, t)}{\partial t} \right]^2 dx \quad (3-3)$$

and the kinetic energy of the RMA is

$$T_R(t) = \frac{1}{2} M \left[\frac{\partial w(L, t)}{\partial t} \right]^2 + \frac{1}{2} M_a [\dot{w}_a(t)]^2 \quad (3-4)$$

Hence, the total kinetic energy of the system $T(t)$ is equal to $T_b(t) + T_R(t)$.

$$T(t) = \frac{1}{2} \int_0^L \rho A \left[\frac{\partial w(x, t)}{\partial t} \right]^2 dx + \frac{1}{2} M \left[\frac{\partial w(L, t)}{\partial t} \right]^2 + \frac{1}{2} M_a [\dot{w}_a(t)]^2 \quad (3-5)$$

Substituting the solution Eq. 3-1 into the above equation and taking the term ρAL out of the summation notations, we have

$$T(t) = \frac{1}{2} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \rho AL \left[\int_0^L \phi_n(x) \phi_m(x) (dx/L) + \alpha(1 + \mu a_n a_m) \phi_n(L) \phi_m(L) \right] \dot{q}_n(t) \dot{q}_m(t) \quad (3-6)$$

where the over-bar denotes the dimensionless form of the mode shapes.

Examining Eq. 3-6, we can see that the term inside the summation signs satisfies the orthogonality derived in Section 2.3. We know that only in the case of $n = m$ the term inside the bracket is equal to unity, otherwise, the term vanishes.

$$\int_0^L \phi_n(x) \phi_m(x) (dx/L) + \alpha(1 + \mu a_n a_m) \phi_n(L) \phi_m(L) = \delta_{mn} \quad (2-59)$$

Therefore, Eq. 3-6 can be simplified as

$$T(t) = \frac{1}{2} \sum_{n=1}^{\infty} \dot{q}_n^2(t) \quad (3-7)$$

The potential energy of the beam is given by the following integral

$$V_b(t) = \frac{1}{2} \int_0^L EI \left[\frac{\partial^2 w(x,t)}{\partial x^2} \right]^2 dx \quad (3-8)$$

and the potential energy of the RMA is

$$V_R(t) = \frac{1}{2} k [w(L,t) - w_a(t)]^2 \quad (3-9)$$

Hence, the total potential energy of the system $V(t)$ is equal to $V_b(t) + V_R(t)$.

$$V(t) = \frac{1}{2} \int_0^L EI \left[\frac{\partial^2 w(x,t)}{\partial x^2} \right]^2 \left(\frac{dx}{L} \right) + \frac{1}{2} k [w(L,t) - w_a(t)]^2 \quad (3-10)$$

Substituting the solution Eq. 3-1 into the above equation, we obtain

$$V(t) = \frac{1}{2} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \left[\int_0^L EI \phi_n''(x) \phi_m''(x) dx + k(1 - a_n)(1 - a_m) \phi_n(L) \phi_m(L) \right] q_n(t) q_m(t) \quad (3-11)$$

With Eqs. 2-54 and 2-26, Eq. 3-11 can be written as

$$V(t) = \frac{1}{2} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \omega_n^2 \rho A L \left[\int_0^L \phi_n(x) \phi_m(x) \left(\frac{dx}{L} \right) + \alpha(1 + \mu a_n a_m) \phi_n(L) \phi_m(L) \right] q_n(t) q_m(t) \quad (3-12)$$

Again, we use the orthogonality Eq. 2-59 to simplify Eq. 3-12 to

$$V(t) = \frac{1}{2} \sum_{n=1}^{\infty} \omega_n^2 q_n^2(t) \quad (3-13)$$

Substituting the total kinetic and potential energies, $T(t)$ and $V(t)$ in Eqs. 3-7 and 3-13, into the Lagrange's equation 3-2 and performing the differentiations, we can have a set of infinite number of independent differential equations, which are

$$\ddot{q}_n(t) + \omega_n^2 q_n(t) = Q_n(t) \quad n = 1, \dots, \infty \quad (3-14)$$

where $Q_n(t)$ are the nonconservative forces of the system.

In this section we consider only the external forces, $f(x, t)$ and $F(t)\delta(x - L)$, (no dampnig force) as the nonconservative forces of the system. The work done on the system by the external forces during a virtual displacement is

$$\begin{aligned} \Delta W(t) &= \int_0^L f(x, t) \Delta w(x, t) dx + F(t) \delta(x - L) \Delta w(x, t) \\ &= \sum_{n=1}^{\infty} \left[\int_0^L f(x, t) \phi_n(x) dx + F(t) \phi_n(L) \right] \Delta q_n(t) \end{aligned} \quad (3-15)$$

Also, the work done on the system by the generalized nonconservative forces $(Q_e)_n(t)$, is

$$\Delta W(t) = \sum_{n=1}^{\infty} (Q_e)_n(t) \Delta q_n(t) \quad (3-16)$$

where the subscript e denotes the external forces. By comparing Eq. 3-15 to 3-16, the generalized nonconservative forces are obtained as

$$(Q_e)_n(t) = \int_0^L f(x, t) \phi_n(x) dx + F(t) \phi_n(L) \quad n = 1, \dots, \infty \quad (3-17)$$

Taking the Laplace transform of Eq. 3-14 with zero initial conditions, $q_n(0) = 0$ and $\dot{q}_n(0) = 0$, we can get the solution of Eq. 3-14 as

$$q_n(t) = \frac{1}{\omega_n} \int_0^t (Q_e)_n(t) \sin \omega_n(t - \tau) d\tau \quad n = 1, \dots, \infty \quad (3-18)$$

Eqs. 2-48, 3-1 and 3-18 are the complete solutions of the undamped system with the external forces, $f(x, t)$ and $F(t)\delta(x - L)$.

Forced Response of the System with Viscous Damping in RMA

In this section we will study the damped response of the system subjected to the external forces. A dash-pot is introduced into the RMA, that is to say, the damping inside the RMA is not equal to zero (Fig. 3.1). The damping forces mainly due to the friction of the bearing are applied to both masses, M and M_a , with the same magnitude but opposite directions, which is modelled as

$$|F_d(t)| = c|\dot{w}(L, t) - \dot{w}_a(t)| \quad (3-19)$$

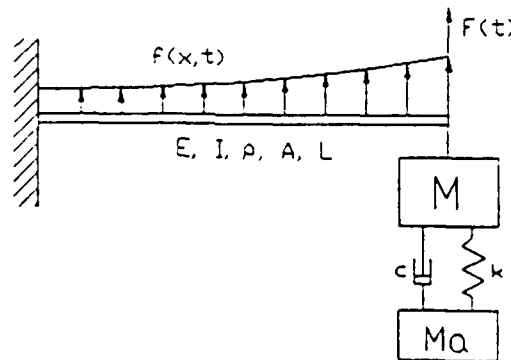


Figure 3.1 System with Dash-Pot in RMA

To solve for the damped response of the system, we use the same approach as we did in the preceding section. Assume the solution is in the form of eigenfunction expansion:

$$\begin{aligned}
 w(x, t) &= \sum_{n=1}^{\infty} \phi_n(x) q_n(t) \\
 w_a(t) &= \sum_{n=1}^{\infty} a_n \phi_n(L) q_n(t)
 \end{aligned} \tag{3-1}$$

where $\phi_n(x)$ and $a_n \phi_n(L)$ are the mode shapes and the $q_n(t)$ are the generalized coordinates.

Based on the approach of energy consideration, the Lagrange's equation is the differential equation, which is expressed in terms of the generalized coordinates as

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_n} \right) - \frac{\partial T}{\partial q_n} + \frac{\partial V}{\partial q_n} = Q_n(t) \quad n = 1, \dots, \infty \tag{3-2}$$

Because the derivation of the differential equation is the same as the one in the preceding section except the for generalized nonconservative forces, we just rewrite the result here.

The total kinetic and potential energies are listed below respectively.

$$T(t) = \frac{1}{2} \sum_{n=1}^{\infty} \dot{q}_n^2(t) \tag{3-7}$$

$$V(t) = \frac{1}{2} \sum_{n=1}^{\infty} \omega_n^2 q_n^2(t) \tag{3-13}$$

The generalized nonconservative forces subjected to the external forces $f(x, t)$ and $F(t)\delta(x - L)$ are

$$(Q_e)_n(t) = \int_0^L f(x, t) \phi_n(x) dx + F(t) \phi_n(L) \tag{3-17}$$

The damping effect of the dash-pot in the RMA is considered as the non-conservative force. The work done on the system by the damping force during a virtual displacement, $\Delta w(L, t)$ and $\Delta w_a(t)$, is

$$\Delta W(t) = -c \left[\frac{\partial w}{\partial t}(L, t) - \dot{w}_a(t) \right] [\Delta w(L, t) - \Delta w_a(t)] \quad (3-20)$$

Substituting Eq. 3-1 into 3-20, we have

$$\Delta W(t) = - \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} [c(1-a_n)(1-a_m)\phi_n(L)\phi_m(L)\dot{q}_m(t)] \Delta q_n(t) \quad (3-21)$$

The work done on the system by the generalized nonconservative forces $(Q_d)_n(t)$ is

$$\Delta W(t) = \sum_{n=1}^{\infty} (Q_d)_n(t) \Delta q_n(t) \quad (3-22)$$

Comparing Eq. 3-22 to 3-21 we can see that

$$(Q_d)_n(t) = - \sum_{m=1}^{\infty} c(1-a_n)(1-a_m)\phi_n(L)\phi_m(L)\dot{q}_m(t) \quad n = 1, \dots, \infty \quad (3-23)$$

Then the total generalized nonconservative forces are

$$\begin{aligned} Q_n(t) &= (Q_e)_n(t) + (Q_d)_n(t) \\ &= \int_0^L f(x, t)\phi_n(x)dx + F(t)\phi_n(L) - \\ &\quad \sum_{m=1}^{\infty} c(1-a_n)(1-a_m)\phi_n(L)\phi_m(L)\dot{q}_m(t) \quad n = 1, \dots, \infty \end{aligned} \quad (3-24)$$

Substituting Eqs. 3-7, 3-13, and 3-24 into the Lagrange's equation EQ. 3-2, we have

$$\ddot{q}_n(t) + \omega_n^2 q_n(t) = Q_n(t) = Q_{en}(t) + Q_{dn}(t) \quad n = 1, \dots, \infty \quad (3-25)$$

Because the $Q_{dn}(t)$ is a function of \dot{q}_m ($m = 1, \dots, \infty$), we can move $(Q_d)_n(t)$ to the left of the equal sign of Eq. 3-25 to get

$$\ddot{q}_n(t) + (Q_d)_n(t) + \omega_n^2 q_n(t) = (Q_e)_n \quad n = 1, \dots, \infty \quad (3-26)$$

Let us denote the column matrix of the generalized coordinates by $q(t)$. The above equation can be written as

$$I\ddot{q}(t) + C\dot{q}(t) + \Omega q(t) = Q_e(t) \quad (3-27)$$

where I is the identity matrix. And

$$\begin{aligned} q(t) &= [q_1(t), q_2(t), \dots, q_n(t)]^T \\ \Omega &= \text{diag}[\omega_n^2] \\ C &= [c_{mn}] \quad c_{mn} = c(1 - a_n)(1 - a_m)\phi_n(L)\phi_m(L) \\ Q_e(t) &= [(Q_e)_1(t), (Q_e)_2(t), \dots, (Q_e)_n(t)]^T \end{aligned}$$

In Eq. 3-27 we also have a set of infinite number of differential equations. Unfortunately, these are coupled equations. The closed-form solution of these equations could not be obtained. Instead, the numerical software package needs to be used to provide the numerical solution of the equation.

For numerical calculations, the system has to be truncated. Therefore, the infinite-degree-of-freedom system becomes an N -degree-of-freedom system. We can have more accurate answers by including more terms in the series.

After solving the generalized coordinates, $q_n(t)$, the displacements at any position along the beam and the mass, M_a , can be calculated from Eq. 3-1. If

we want to have the differential equation of the system in terms of the physical coordinates instead of the generalized coordinates. A transformation matrix defined below can do the work. Let us define a matrix, \mathbf{G} , as

$$\mathbf{G} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{bmatrix} \quad (3-28)$$

where

$$a_{mn} = \phi_n(x = \frac{m}{N}L)$$

and N is the truncated size of the infinite-degree-of-freedom system. Therefore, the relation between the physical and generalized coordinates is

$$\mathbf{w}(t) = \mathbf{G}\mathbf{q}(t) \quad (3-29)$$

where

$$\begin{aligned} \mathbf{w}(t) &= [w(\frac{L}{N}, t), w(\frac{2}{N}L, t), \dots, w(L, t)]^T \\ \mathbf{q}(t) &= [q_1(t), q_2(t), \dots, q_N(t)]^T \end{aligned}$$

The differential equation in terms of the physical coordinates defined above is

$$\ddot{\mathbf{w}}(t) + \mathbf{G}\mathbf{C}\mathbf{G}^{-1}\dot{\mathbf{w}}(t) + \mathbf{G}\mathbf{\Omega}\mathbf{G}^{-1}\mathbf{w}(t) = \mathbf{G}\mathbf{Q}_e(t) \quad (3-30)$$

Eq. 3-30 can be reduced to the undamped system by setting the damping $c = 0$, which is

$$\ddot{\mathbf{w}}(t) + \mathbf{G}\mathbf{\Omega}\mathbf{G}^{-1}\mathbf{w}(t) = \mathbf{G}\mathbf{Q}_e(t) \quad (3-31)$$

Nondimensionalization and Case Study

In order to facilitate the numerical calculations, the differential equation of the system is usually nondimensionalized and expressed in the state-space form. Let us define the following dimensionless parameters:

$$\begin{aligned}\bar{w}(t) &= \frac{w(t)}{\frac{|F(t)|}{\rho AL \omega_b^2}} & \dot{\bar{w}}(t) &= \frac{\dot{w}(t)}{\frac{|F(t)|}{\rho AL \omega_b}} & \ddot{\bar{w}}(t) &= \frac{\ddot{w}(t)}{\frac{|F(t)|}{\rho AL}} \\ \bar{F}(t) &= \frac{F(t)}{|F(t)|} & \bar{f}(x, t) &= \frac{f(x, t)}{|F(t)|}\end{aligned}\quad (3-32)$$

where $\omega_b = \sqrt{(EI)/(\rho AL^4)}$.

And the dimensionless mode shapes are defined as

$$\bar{\phi}_n(x) = \bar{C}_n h_n(x) \quad \bar{\phi}_a = a_n \bar{C}_n h_n(L) \quad (3-33)$$

where

$$\begin{aligned}h_n(x) &= (\cosh \lambda_n x - \cos \lambda_n x) - \frac{\cosh \lambda_n L + \cos \lambda_n L}{\sinh \lambda_n L + \sin \lambda_n L} (\sinh \lambda_n x - \sin \lambda_n x) \\ \bar{C}_n &= \sqrt{\rho AL} C_n = \left[\int_0^L h_n^2(x) (dx/L) + \alpha(1 + \mu a_n^2) h_n^2(L) \right]^{-1/2} \\ a_n &= \frac{f_a^2}{f_a^2 - (\lambda L)^4}\end{aligned}$$

then, the transformation matrix, \mathbf{G} , is dimensionless. For consistency we denote it as $\bar{\mathbf{G}}$. The dimensionless natural frequencies are defined as, $\omega_n/\omega_b = (\lambda_n L)^2$, $n = 1, 2, \dots, \infty$. Hence, the dimensionless matrix, $\bar{\Omega}$, is

$$\bar{\Omega} = \text{diag}[(\lambda_n L)^4] \quad (3-34)$$

Recall the mass ratio parameters defined before are

$$\alpha = \frac{M}{\rho AL} \quad \mu = \frac{M_a}{M} \quad (2-27)$$

If the damping ratio, ζ , is defined as

$$\zeta = \frac{c}{2M_a\omega_b} \quad (3-35)$$

the coefficient matrix, \mathbf{C} , can be arranged in dimensionless form as

$$\bar{\mathbf{C}} = [c_{mn}] \quad c_{mn} = 2\alpha\mu\zeta(1 - a_m)(1 - a_n)\bar{\phi}_m(L)\bar{\phi}_n(L) \quad (3-36)$$

And the dimensionless external force matrix $\bar{\mathbf{Q}}_e$ is

$$\begin{aligned} \bar{\mathbf{Q}}_e &= [(\bar{Q}_e)_n] \\ (\bar{Q}_e)_n &= \int_0^1 \bar{f}(x, t)\bar{\phi}(x)dx + \bar{F}(t)\bar{\phi}(L) \end{aligned} \quad (3-37)$$

With these dimensionless parameters, the differential equation 3-30 is arranged in dimensionless form as

$$\ddot{\bar{\mathbf{w}}}(t) + \bar{\mathbf{G}}\bar{\mathbf{C}}\bar{\mathbf{G}}^{-1}\dot{\bar{\mathbf{w}}}(t) + \bar{\mathbf{G}}\bar{\Omega}\bar{\mathbf{w}}^{-1}(t)\bar{\mathbf{w}}(t) = \bar{\mathbf{G}}\bar{\mathbf{Q}}_e(t) \quad (3-38)$$

Let us define the state vector, $\mathbf{x}(t)$, as

$$\mathbf{x}(t) = [w(\frac{L}{N}, t), w(\frac{2}{N}L, t), \dots, w(L, t), \dot{w}(\frac{L}{N}, t), \dot{w}(\frac{2}{N}L, t), \dots, \dot{w}(L, t)]^T \quad (3-39)$$

The state-space form of the differential equation in terms of the physical coordinates $\mathbf{x}(t)$ is

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{P}(t) \quad (3-40)$$

where

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \frac{\partial \mathbf{x}}{\partial t} \\ \mathbf{A} &= \begin{bmatrix} \mathbf{I} & 0 \\ \bar{\mathbf{G}}\bar{\Omega}\bar{\mathbf{w}}^{-1} & \bar{\mathbf{G}}\bar{\mathbf{C}}\bar{\mathbf{G}}^{-1} \end{bmatrix} \\ \mathbf{P}(t) &= \begin{bmatrix} 0 \\ \bar{\mathbf{G}}\bar{\mathbf{Q}}_e \end{bmatrix} \end{aligned}$$

Case Study

Two computer programs are written for the numerical calculations. The first one written in FORTRAN computer code is for calculating the numerical values of elements of the coefficient matrices of the state-space equation 3-40. The second one is for MATRIX_X input which is used to calculate the response of the system subjected the external forces. (See Appendix.)

The following examples are calculated.

case	α	μ	f_a	ζ
1	0.25	1.0	0	0
2	0.25	1.0	1.8	0
3	0.25	1.0	1.8	0.5

The frequency and impulse response of the displacement $\bar{w}(L, t)$ is plotted in Figs. 3.2, and 3.3.

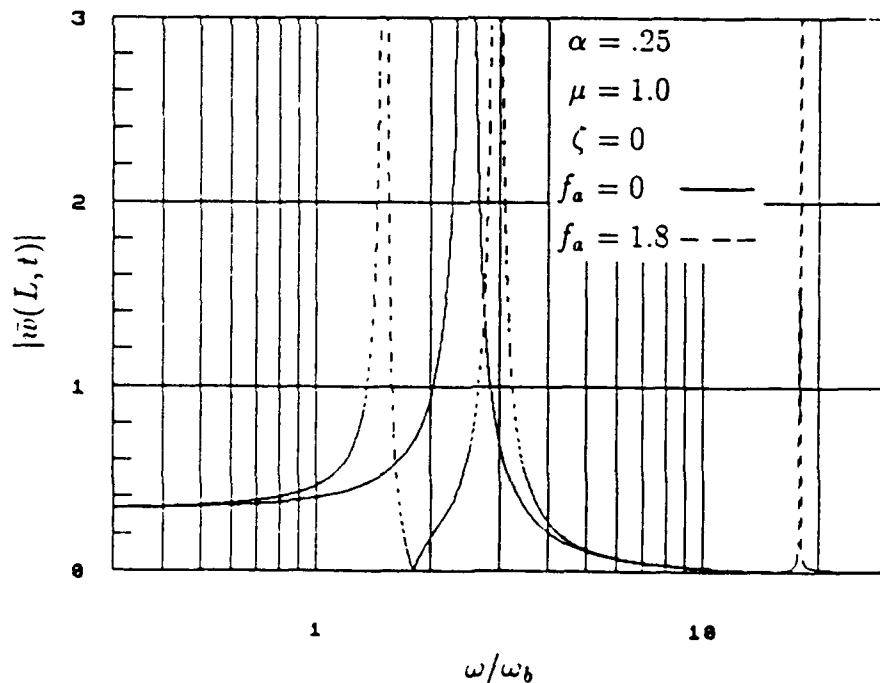


Figure 3.2 Frequency Response of the Undamped System

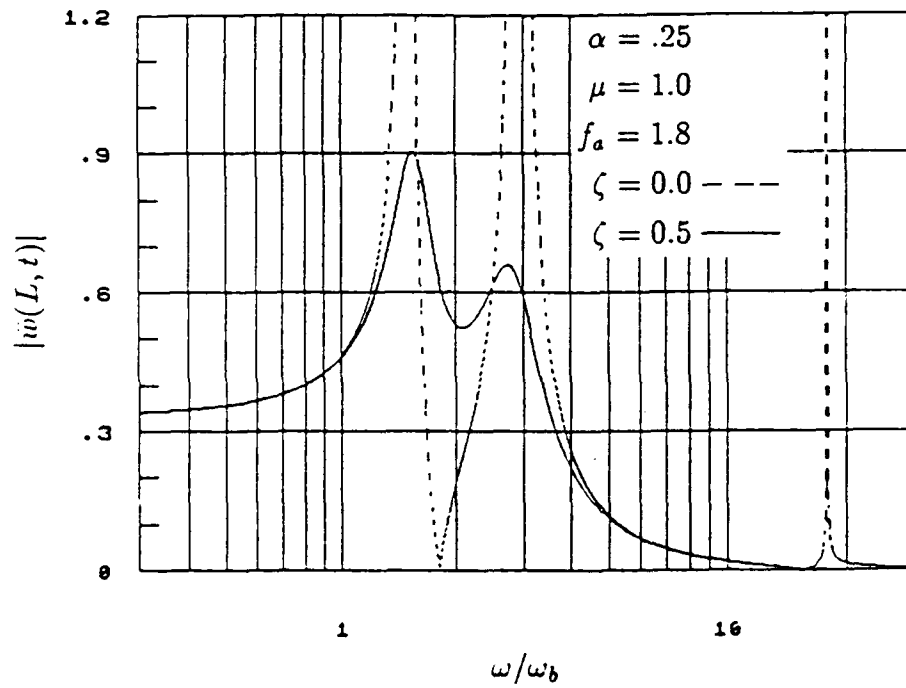


Figure 3.3 Frequency Response of the System without and with Damping

CHAPTER 4

Active Vibration Absorber Design

State Differential Equation with Control Input

The mathematical model of the beam-RMA system has been derived in Chapter 3. In this section we analyze the effect of control input to the system and derive the differential equation for the system which is called the open-loop control system. The derivation is based on the approach of energy consideration which is the same as what we did in Chapter 3. The purpose of incorporating the RMA to the beam is to generate the control force which, by its nature, applies to both masses M and M_a with the same magnitude but opposite directions. (Fig. 4.1)

The RMA actually is a DC motor from which force is generated by electromagnetic effect. Consider a charge q moving with velocity V in a magnetic field of intensity B experiencing a force u given by the vector cross-product $\vec{u} = q\vec{V} \times \vec{B}$. Because the current i is equal to dq/dt , the force u has the magnitude

$$u(t) = Bli(t) \quad (4-1)$$

if the direction of current is at right angle of magnetic field. Eq. 4-1 is the law of motors. Note that the control force is proportional to the current input and the direction of the force also depends on the direction of the current.

Consider the model in Fig. 4.1, the derivation of state differential equation is the same as the procedure has gone through in Chapter 3 except that we need to add the effect of the control force which is nonconservative. From the approach

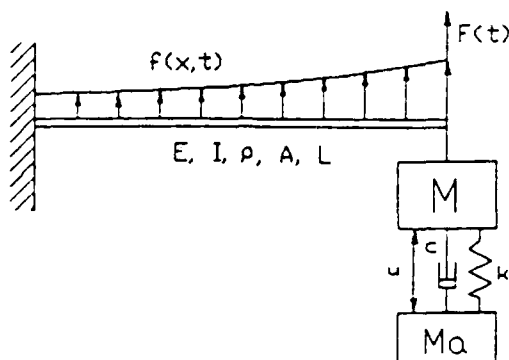


Figure 4.1 Beam-RMA System with Control Force Input

of energy consideration, the virtual work done on the system by the control force $u(t)$ during a virtual displacement $\Delta w(L, t)$ and $\Delta w_a(t)$ is

$$\Delta W(t) = u(t)[\Delta w(L, t) - \Delta w_a(t)] \quad (4-2)$$

By the assumption of eigenfunction expansion, the solution can be expressed as

$$w(x, t) = \sum_{n=1}^{\infty} \phi_n(x) q_n(t), \quad w_a(t) = \sum_{n=1}^{\infty} a_n \phi_n(L) q_n(t) \quad (4-3)$$

By substituting Eq. 4-3 into 4-2, Eq. 4-2 can be written as

$$\Delta W(t) = \sum_{n=1}^{\infty} u(t)(1 - a_n \phi_n(L)) \Delta q_n(t) \quad (4-4)$$

Also, the work done on the system by the generalized nonconservative forces $Q_{cn}(t)$ is

$$\Delta W(t) = \sum_{n=1}^{\infty} Q_{cn}(t) \Delta q_n(t) \quad (4-5)$$

Comparing Eq. 4-4 to 4-5, we obtain the generalized nonconservative forces due to the control input $u(t)$ as

$$Q_{cn}(t) = (1 - a_n)\phi_n(L)u(t) \quad n = 1, \dots, \infty \quad (4-6)$$

Arranging Eq. 4-6 into the matrix form and adding it to the system equation 3-27 in Chapter 3, we have

$$I\ddot{\mathbf{q}}(t) + C\dot{\mathbf{q}}(t) + \Omega\mathbf{q}(t) = \mathbf{Q}_e(t) + \mathbf{Q}_c(t) \quad (4-7)$$

where

$$\mathbf{Q}_c(t) = [Q_{c1}(t), Q_{c2}(t), \dots, Q_{cn}(t)]^T.$$

Because the generalized coordinates $q(t)$ can not be measured for feedback, we need to transform Eq. 4-7 into an equation in terms of the physical coordinates $\mathbf{x}(t)$ which can be sensed by the accelerometers. This has already been done in Chapter 3. We just quote the result and apply it to Eq. 4-7 to have

$$\ddot{\mathbf{w}}(t) + \mathbf{GCG}^{-1}\dot{\mathbf{w}}(t) + \mathbf{G}\Omega\mathbf{G}^{-1}\mathbf{w}(t) = \mathbf{GQ}_e(t) + \mathbf{GQ}_c(t) \quad (4-8)$$

Again, through the same procedure to nondimensionlize Eq. 4-8, we can have

$$\ddot{\bar{\mathbf{w}}}(t) + \bar{\mathbf{G}}\bar{\mathbf{C}}\bar{\mathbf{G}}^{-1}\dot{\bar{\mathbf{w}}}(t) + \bar{\mathbf{G}}\bar{\Omega}\bar{\mathbf{G}}^{-1}\bar{\mathbf{w}}(t) = \bar{\mathbf{G}}\bar{\mathbf{Q}}_e(t) + \bar{\mathbf{G}}\bar{\mathbf{Q}}_c(t) \quad (4-9)$$

where the elements of $\bar{\mathbf{Q}}_c(t)$ are

$$(1 - a_n)\bar{\phi}_n(L)\frac{u(t)}{|F(t)|}$$

The state-space form of Eq. 4-9 is

$$\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{B}\bar{u}(t) + \mathbf{P}(t) \quad (4-10)$$

where $\mathbf{x}(t)$, \mathbf{A} , and $\mathbf{P}(t)$ are described in Section 3.3, and

$$\mathbf{B} = \begin{bmatrix} 0 \\ \bar{\mathbf{G}}\bar{\mathbf{Q}}_c' \end{bmatrix} \quad \bar{u}(t) = \frac{u(t)}{|F(t)|}$$

$$\bar{\mathbf{Q}}_c' = [(1 - a_1)\bar{\phi}_1(L), (1 - a_2)\bar{\phi}_2(L), \dots, (1 - a_n)\bar{\phi}_n(L)]^T.$$

Eq. 4-10 is a linear time-invariant system because the coefficient matrices \mathbf{A} and \mathbf{B} are constant. We will use it as the mathematical model to derive the feedback control law.

Stability and Controllability

In this section we are interested in the overall time behavior of the differential system to see whether or not the solutions of the state differential equation tend to grow infinitely as $t \rightarrow \infty$. This is called the stability of the system.

For solving the control problem, it is also important to know whether or not the system has the property that it may be steered from any given state to any other given state. This leads to the concept of controllability of the system.

Before starting to design the control system, we should discuss both stability and controllability to see how stable the system is and make sure that it is controllable.

The state differential equation for the system with control force input has been derived in preceding section which is

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\bar{u}(t) + \mathbf{P}(t) \quad (4-11)$$

which is a linear and time-invariant system with dimension $2N$. From control theory we know that the stability of the system 4-11 depends on the coefficient

matrix A . The system 4-11 is said to be exponentially stable if and only if all the eigenvalues of the matrix A have strictly negative real parts. That is to say, all the poles of the open-loop control system have to be within the left-hand side of the complex plane. If this condition is valid for the system, we can be sure that the solutions of Eq. 4-11 with finite input $P(t)$ will not grow with bound.

Setting $\bar{u}(t)$ and $P(t)$ equal to zero in Eq. 4-11, and taking the Laplace transform of the Eq. 4-11, we have

$$sX = AX$$

$$[sI - A]X = 0 \quad (4-12)$$

In order to have nontrivial solution for Eq. 4-12, the determinant of matrix $sI - A$ must be equal to zero

$$|sI - A| = 0 \quad (4-13)$$

The eigenvalues of the matrix A can be determined by solving the Eq. 4-13. Expanding the determinant gives a $2N^{th}$ -order polynomial of variable s . The roots of the polynomial are the eigenvalues of the open-loop control system which are denoted by $\lambda_1, \lambda_2, \dots, \lambda_{2N}$.

From control theory, the controllability depends on the coefficient matrices A and B . The system 4-11 is said to be completely controllable if and only if the rank of the controllability matrix H

$$H = [B, AB, A^2B, \dots, A^{2N-1}B] \quad (4-14)$$

is $2N$ which means that the determinant of matrix \mathbf{H} is not equal to zero. If this condition is valid, a control system for steering from a given state to any other given state is possible to find.

Case Study

Throughout Chapters 4 and 5, the discussion and design of the feedback control loop are based on the system model including the first two vibration modes. Then the state differential equation is a 4-dimensional system. The state vector chosen is

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \ddot{w}(.5L, t) \\ \ddot{w}(L, t) \\ \dot{w}(.5L, t) \\ \dot{w}(L, t) \end{bmatrix} \quad (4-15)$$

The numerical values for the system parameters are assigned as

$$\alpha = 0.25 \quad \mu = 1.0 \quad f_a = 1.8 \quad \zeta = 0.5$$

$$f(x, t) = 0 \quad \frac{F(t)}{|F(t)|} = e^{j\bar{\omega}t} \quad \bar{\omega} = \frac{\omega}{\omega_b} \quad (4-16)$$

From the output of the FORTRAN computer program (listed in Appendix), we have the coefficient matrices

$$\bar{\mathbf{G}} = \begin{bmatrix} .1753 & .4341 \\ .5525 & 1.310 \end{bmatrix} \quad \bar{\mathbf{\Omega}} = \begin{bmatrix} 2.270 & 0 \\ 0 & 8.739 \end{bmatrix} \quad \bar{\mathbf{C}} = \begin{bmatrix} .0835 & -.1346 \\ -.1346 & .2168 \end{bmatrix}$$

$$\bar{\mathbf{Q}}'_e = \begin{bmatrix} .5525 \\ 1.310 \end{bmatrix} \quad \bar{\mathbf{Q}}'_c = \begin{bmatrix} -1.292 \\ 2.082 \end{bmatrix} \quad (4-17)$$

Stability

The state differential equation of the open-loop control system without damping and external forces is calculated from Eqs. 4.9 and 4.10 as

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -154.42 & 48.275 & 0 & 0 \\ -459.15 & 143.41 & 0 & 0 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ .677 \\ 2.014 \end{bmatrix} \bar{u}(t) \quad (4-18)$$

The stability of Eq. 4-18 depends on the eigenvalues of matrix A which are calculated as

$$\pm 1.507i, \pm 2.956i$$

These four poles are located on the imaginary axis and the system is neutrally stable which means that the response neither dies out nor grow to infinity. This is what we expected because there is no damping in the system.

The state differential equation of the open-loop system with damping is

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -154.42 & 48.275 & -47.235 & 15.38 \\ -459.15 & 143.41 & -140.42 & 54.73 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ .677 \\ 2.014 \end{bmatrix} \bar{u}(t) \quad (4-19)$$

The poles of the open-loop system are

$$-.1083 \pm 2.950i, \quad -.0418 \pm 1.5082i \quad (4-19a)$$

The damped frequencies and damping ratios for the first two modes are

$$\begin{aligned} (\bar{\omega}_d)_1 &= 2.952, & (\bar{\omega}_d)_2 &= 1.509 \\ \zeta_1 &= .0367, & \zeta_2 &= .0277 \end{aligned}$$

The poles of the damped system are within the left-hand plane and it is exponentially stable. We see that the damping ratios for the two modes are very low. This explains the resonant phenomenon of the frequency response.

Controllability

The controllability matrix of the undamped system is

$$H = \begin{bmatrix} 0 & .667 & 0 & -7.317 \\ 0 & 2.014 & 0 & -22.014 \\ .677 & 0 & -7.317 & 0 \\ 2.014 & 0 & -22.014 & 0 \end{bmatrix}$$

which can be easily shown that it has the rank 4 and the system is completely controllable.

The controllability matrix of the damped system is

$$H = \begin{bmatrix} 0 & .667 & -1.003 & -5.539 \\ 0 & 2.014 & -2.964 & -16.756 \\ .677 & -1.003 & -5.539 & 15.658 \\ 2.014 & -2.964 & -16.756 & 46.799 \end{bmatrix}$$

of which the rank 4 may be shown to be 4 and the system is completely controllable.

Performance Improvement by State

Feedback Control

From preceding section, we know that the open-loop control system is slightly damped because the poles of the system are close to the imaginary axis. This explains why the resonant phenomenon occurs when the system is subjected to a sinusoidal input. In this section we use state feedback control as the tool to suppress the resonant vibration of the system by improving its performance, which also means that we can push the poles far to the left of the complex plane by feedback control. This is called the pole-placement technique.

The closed-loop system is made up of sensors, signal amplifiers, and the RMA (Fig. 4.2). The accelerations are measured by the accelerometers and integrated twice to yield the velocities and displacements for feedback. The amplifiers are modelled as a set of constant gains. The RMA is the actuator of the control system which generates the control force proportional to the current input.

From the case study of preceding section, we saw that our system is completely controllable which guarantees that a control law can be found. By adjusting the gain values, the eigenvalues of the closed-loop system can be changed. The main effort in this section is to determine the gain matrix to have the preferable frequency response.

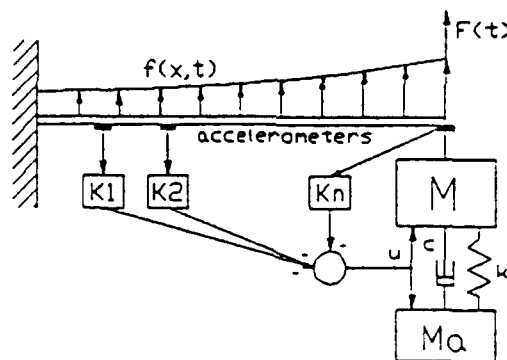


Figure 4.2 State Feedback Control System

The state differential equation has been derived in the first section of this chapter which is a linear and time-invariant system.

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\bar{\mathbf{u}}(t) + \mathbf{P}(t) \quad (4-20)$$

Suppose that the complete state can be accurately measured at all times, the control law implemented in Fig. 4.2 is linear which has the form

$$\bar{u}(t) = -Kx(t) \quad (4-21)$$

where K is the feedback gain matrix.

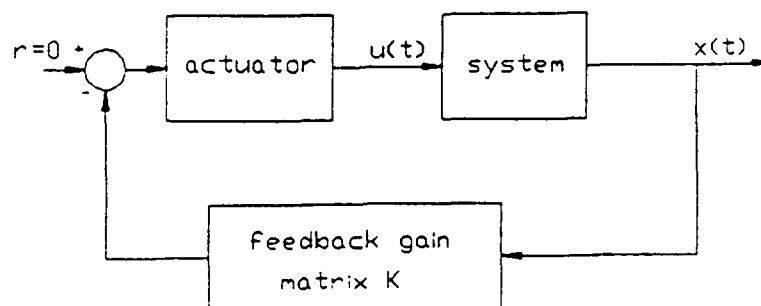


Figure 4.3 Block Diagram of the Feedback Control System

If this control law is connected to the system 4-20, the closed-loop system (Fig. 4.3) is described by the following state differential equation

$$\dot{x}(t) = (A - BK)x(t) + P(t) \quad (4-22)$$

The dynamics of the closed-loop system depend on the eigenvalues, which is referred as the closed-loop poles of the system. The poles are determined from the matrix $A - BK$. By choosing suitable constant gain matrix K , we can move the closed-loop poles to left in the complex plane to have required system response. Because our system has single-input control force u , the gain matrix is uniquely found for a given set of closed-loop poles.

From the eigenvalue problem analysis, we set the external forces $P(t) = 0$ in Eq. 4-22 and take the Laplace transform of this equation to have

$$sX = (A - BK)X$$

$$[s\mathbf{I} - (\mathbf{A} - \mathbf{BK})]\mathbf{X} = 0 \quad (4-23)$$

The determinant of the the matrix $s\mathbf{I} - (\mathbf{A} - \mathbf{BK})$ must be equal to zero to have the nontrivial solution.

$$|s\mathbf{I} - (\mathbf{A} - \mathbf{BK})| = 0 \quad (4-24)$$

If the system is truncated to have N modes, then the characteristic polynomial obtained by expanding the determinant of Eq. 4-24 is a $2N^{\text{th}}$ -order polynomial. It has $2N$ poles which can be located at any positions by assinging siutable gains in the \mathbf{K} matrix. Let the closed-loop poles are $\lambda_1, \lambda_2, \dots, \lambda_{2N}$. The corresponding polynomial in factor form is

$$(s - \lambda_1)(s - \lambda_2) \dots (s - \lambda_{2N}) \quad (4-25)$$

Comparing the coefficients of both polynomials obtained from Eqs. 4-24 and 4-25, we have $2N$ equations from which the $2N$ unknown gains, k_1, k_2, \dots, k_{2N} can be obtained. Following is a study of feedback control system.

Case Study

We continue the study in Section 4.2 to illustrate the performance improvement by the state feedback control system. To suppress the resonance of the first two modes of the combined system, a two-position feedback loop is a good model to work with (Fig. 4.4). We measure the states at positions $x = 0.5L$ and $x = L$ for feedback. The state differential equation of the open-loop system is given by Eq. 4-19

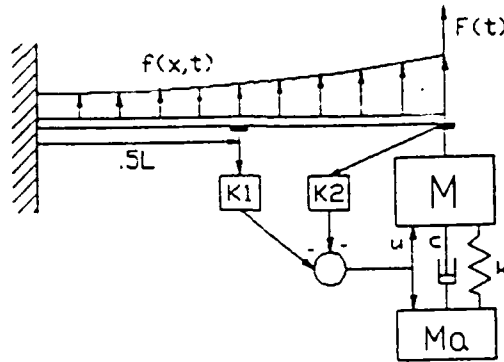


Figure 4.4 Two-Position Feedback Control System

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -154.42 & 48.275 & -47.235 & 15.38 \\ -459.15 & 143.41 & -140.42 & 45.73 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ .677 \\ 2.014 \end{bmatrix} \bar{u}(t) \quad (4-19)$$

The poles of the open-loop system are

$$-.1083 \pm 2.950i, \quad -.0418 \pm 1.5082i \quad (4-19a)$$

Let us consider the control law

$$\bar{u}(t) = -[k_1, k_2, k_3, k_4]\mathbf{x}(t) \quad (3-26)$$

Suppose we want the closed-loop poles, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, to be located at $-1.5 \pm .2i$ and $-.5 \pm .1i$, then the control law is calculated as

$$\bar{u}(t) = -[-219.8, 71.22, -33.84, 12.62]\mathbf{x}(t) \quad (4-27)$$

Substituting Eq. 4-27 into Eq. 4-19 and adding a concentrated force $F(t)\delta(x-L)$, we have

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -6028 & 1963 & -5133 & 1718 \\ -17933 & 5840 & -15270 & 5110 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ .665 \\ 2.021 \end{bmatrix} \bar{F}(t) \quad (4-28)$$

For sinusoidal input, $\bar{F}(t) = e^{j\omega t}$. For impulse input $\bar{F}(t) = \delta(t)$. The frequency response and impulse response of the displacement $\bar{w}(L, t)$ of the system without and with state feedback control system are shown in Figs. 4.5 and 4.6, respectively.

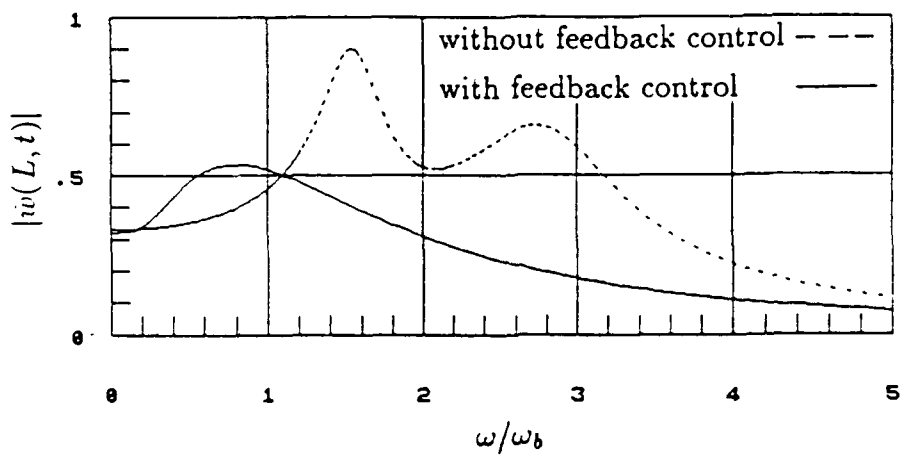


Figure 4.5 Frequency Response of the System without and with State Feedback Control System

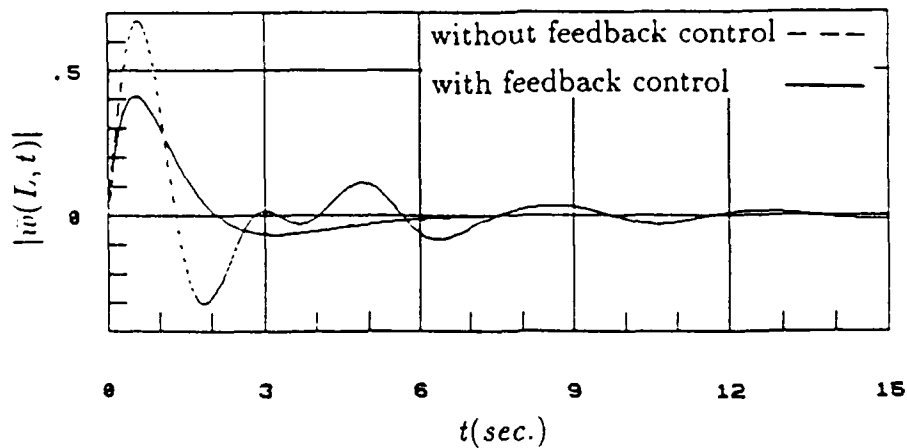


Figure 4.6 Impulse Response of the System without and with State Feedback Control System

Optimal Linear Quadratic Regulator

Design

In preceding section, we saw that the performance of the system can be improved by state feedback control system. By placing the closed-loop poles to the left of the complex plane, the frequency response of the system can be improved to any degree that we want. However, this is under the assumption that control force is unlimited. In any practical problem the control force must be bounded. This imposes a limit on the distance over which the closed-loop poles can be moved to the left. This consideration leads to the formulation of an optimization problem which takes both the effect of vibration reduction and the magnitude of the control force into the design consideration.

The closed-loop system derived in the preceding section can be expressed in state-space form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\bar{\mathbf{u}}(t) + \mathbf{P}(t) \quad (4-10)$$

which is a linear time-invariant state equation with linear feedback control law

$$\bar{\mathbf{u}}(t) = -\mathbf{K}\mathbf{x}(t) \quad (4-29)$$

The design goal of optimal control system is to find a control law, which under the limitation of the control effect, can reduce the vibration of the beam as much as possible. A performance index has to be defined to have a mathematical statement of the design goal. Based on the performance index, the optimal design is found.

The design goal can be divided into two parts. One is to minimize the vibration, the other is to reduce the control effect. Usually a quadratic integral

form is employed as the index. The index for vibration of the beam is chosen as following

$$\int_0^{\infty} \mathbf{x}^T(t) \mathbf{R}_{xx} \mathbf{x}(t) dt \quad (4-30)$$

where $\mathbf{x}(t)$ is the physical coordinates which are the particular positions for where the vibration need to be suppressed and the \mathbf{R}_{xx} is a nonnegative-definite symmetric weighting matrix.

The index for control effect is

$$\int_0^{\infty} \bar{u}(t) R_{uu} \bar{u}(t) dt \quad (4-31)$$

where the \mathbf{R}_{uu} is a positive-definite symmetric weighting matrix.

Hence, the performance index taking both effects into consideration is

$$\mathbf{J} = \int_0^{\infty} [\mathbf{x}^T(t) \mathbf{R}_{xx} \mathbf{x}(t) + \bar{u}(t) R_{uu} \bar{u}(t)] dt \quad (3-32)$$

If we can find a control force $\bar{u}^*(t)$ such that the performance index \mathbf{J} is minimized. we call

$$\bar{u}^*(t) = -\mathbf{K}^* \mathbf{x}(t) \quad (4-33)$$

the optimal feedback control law with the optimal feedback gain matrix \mathbf{K}^* . The system 4-10 with Eq. 4-33 is called an optimal linear quadratic regulator (LQR).

Case Study

We continue the study in Section 4.3 with the state differential equation

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -154.42 & 48.275 & -47.24 & 15.38 \\ -459.15 & 143.41 & -140.42 & 45.73 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ .677 \\ 2.014 \end{bmatrix} \bar{u}(t) \quad (4-19)$$

If the weighting matrices \mathbf{R}_{xx} and \mathbf{R}_{uu} are chosen as

$$\mathbf{R}_{xx} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{R}_{uu} = [1 * 10^{-3}] \quad (4-34)$$

which means that the design goal is to reduce the vibration of the beam at position of $x = L$ and the number chosen for \mathbf{R}_{uu} depends on the control force limitation.

The optimal feedback control law calculated is

$$\bar{u}^*(t) = -[-51.53, 48.82, 425.3, -138.2]\mathbf{x}(t) \quad (4-35)$$

Substituting Eq. 4-35 into Eq. 4-10, we have the optimal closed-loop system as

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -120 & 15.2 & -335 & 109 \\ -335 & 45.1 & -997 & 324 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ .665 \\ 2.021 \end{bmatrix} \bar{F}(t) \quad (4-36)$$

The frequency response and impulse response of the displacement $\bar{w}(L, t)$ of the system with and without optimal feedback control are shown in Figs. 3.7 and 3.8, respectively.

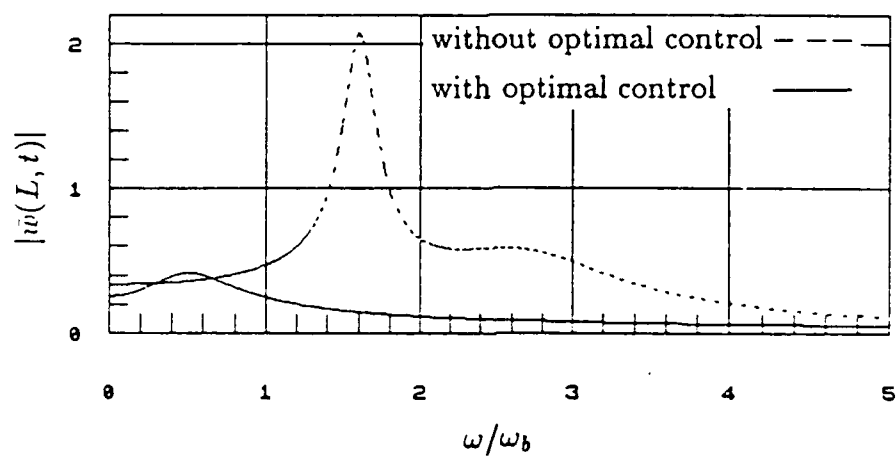


Figure 4.7 Frequency Response of the System without and with Optimal Control System

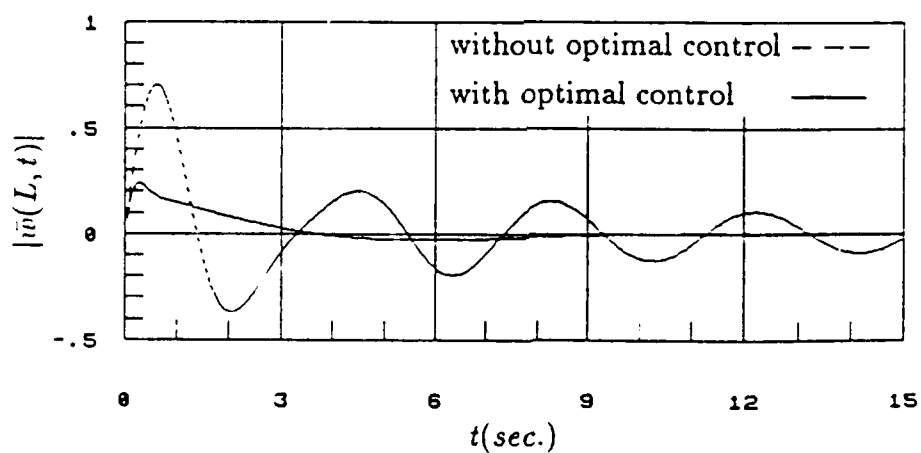


Figure 4.8 Impulse Response of the System without and with Optimal Control System

Sensitivity

As we saw in the preceding study, the designed feedback control system is based on a truncated-mode model which has a discrepancy between the real system. This can be viewed as a kind of parameter changes in the modeling procedure. Besides, the assumptions of modeling procedure also bring this kind effects to the system equation. In this section study whether or not the optimal control system subjected to the parameter changes still works.

A very important property of the feedback control system is its ability to suppress noise and to compensate for parameter changes. We will investigate to what extent the system with optimal LQR design possesses this property. We call this property the sensitivity of the system. Our study is limited to the effects of parameter changes and consider only the steady-state case where the terminal time is infinite.

The optimal feedback control system is

$$\dot{x}(t) = Ax(t) + B\bar{u}(t) + P(t) \quad (4-10)$$

with the optimal control law

$$\bar{u}^*(t) = -K^*x(t) \quad (4-33)$$

From control theory (See Appendix 7, Chapter 3), the sensitivity of the closed-loop system for compensating the parameter changes as compared to an equivalent opne-loop configuration is determined by the return difference matrix $J(s)$

$$J(s) = I + (sI - A)^{-1}BK^* \quad (4-37)$$

If the condition

$$\mathbf{J}^T(-j\omega)\mathbf{W}\mathbf{J}(j\omega) \geq \mathbf{W} \quad (4-38a)$$

for every real input frequency ω is satisfied, where the matrix \mathbf{W} is the weighting matrix of the sensitivity criterion 4-38,

$$\mathbf{W} = \mathbf{K}^{*T}\mathbf{R}_{uu}\mathbf{K}^* \quad (4-38b)$$

we can guarantee that the optimal control system design has the ability to compensate the parameter changes.

Because the matrix $\mathbf{K}^{*T}\mathbf{R}_{uu}\mathbf{K}^*$ is known after the control law has been computed, it is difficult to choose the design parameters \mathbf{R}_{xx} and \mathbf{R}_{uu} such that the condition 4-38 can be achieved. However, under the condition that the poles of the equivalent open-loop controlled system are all within the left-hand plane, the condition 4-38 can be guaranteed.

The equivalent statement of condition 4-38 is the same as that if the condition

$$\mathbf{W} \rightarrow \mathbf{R}_{xx} \quad (4-39)$$

as $\mathbf{R}_{uu} \rightarrow 0$ is satisfied, the sensitivity of the system can be guaranteed.

Case Study

From the case study in Section 4.4, if the weighting matrices \mathbf{R}_{xx} and \mathbf{R}_{uu} are

$$\mathbf{R}_{xx} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{R}_{uu} = [1 \times 10^{-7}]$$

the optimal feedback gain matrix K^* calculated from $MATRIX_X$ is

$$K^* = [-224., 3232, 452.6, -96.3]$$

The weighting matrix W in Eq. 4-38 is

$$W = K^{*T} R_{uu} K^* = \begin{bmatrix} .0005 & -.0228 & -.0004 & -.0001 \\ -.02280 & 1.0142 & .0193 & .0035 \\ -.0004 & .0193 & .0004 & .0001 \\ -.0001 & .0035 & .0001 & .0000 \end{bmatrix}$$

The matrix W is quite close to the limiting matrix R_{xx} . This means that the design of control system based on the truncated model guarantees that it has enough ability to compensate the parameter changes and to suppress the noise.

CHAPTER 5
Estimator Design and Output Feedback
Control System

Reconstructability

Because the linear optimal control law requires the complete state for feedback, we have to reconstruct the complete state when the state measurement is incomplete and inaccurate. In this section we study whether or not the complete state can be reconstructed from the measured state (output). We call this ability the reconstructability of the system. Before designing an estimator, we have to discuss the reconstructability to make sure that such an estimator can be designed.

The state differential equation for the system with output equation is given by the following

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\bar{\mathbf{u}}(t) \\ y(t) &= \mathbf{C}'\mathbf{x}(t)\end{aligned}\tag{5-1}$$

which is a linear and time-invariant system with dimension $2N$. From control theory we know that the reconstructability of the system 5-1 depends on the coefficient matrices \mathbf{A} and \mathbf{C}' . The system 5-1 is said to be completely reconstructable if and only if the rank of the reconstructability matrix \mathbf{H}' ,

$$H' = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{2N-1} \end{bmatrix} \quad (5-2)$$

is $2N$ which means that the determinant of matrix H' is not equal to zero. If this condition is valid, an estimator for reconstructing the complete state from an incompletely measured output is possible to design.

Case Study

We continue the case studied in Chapter 4. The open-loop control system is described by the differential equation

$$\dot{x}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -154.42 & 48.275 & -47.235 & 15.38 \\ -459.15 & 143.41 & -140.42 & 45.73 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \\ .677 \\ 2.014 \end{bmatrix} \bar{u}(t) \quad (5-3)$$

If the output state, that we want to measure, is the displacement $\bar{w}(L, t)$, then the output coefficient matrix C' is

$$C' = [0, 1, 0, 0] \quad (5-4)$$

$$\bar{w}(L, t) = [0, 1, 0, 0]x(t) \quad (5-5)$$

The reconstructability of the system 5-1 depends on the matrix H' , calculated as

$$H' = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -459.153 & 143.412 & -140.42 & 45.73 \\ 686.73 & -220.64 & -247.82 & 74.98 \end{bmatrix}$$

which has the rank 4. This means that the state of system 5-1 is completely constructable. Therefore, the estimator design will be possible.

Full-Order Estimator Design

When the complete state is not available for feedback or there is a lot of noise in measurement, we need to reconstruct the complete state from the partially measured state. As we can see from Chapter 4, the more vibration modes are included in the model, the more accelerators and integrators we will need for the feedback loop. With this in mind, we can imagine that the feedback loop will become very complicated. This situation usually is not feasible. An estimator will be the answer to this problem. Our study here is focused on full-order estimator design which means the estimator can generate the complete state for feedback.

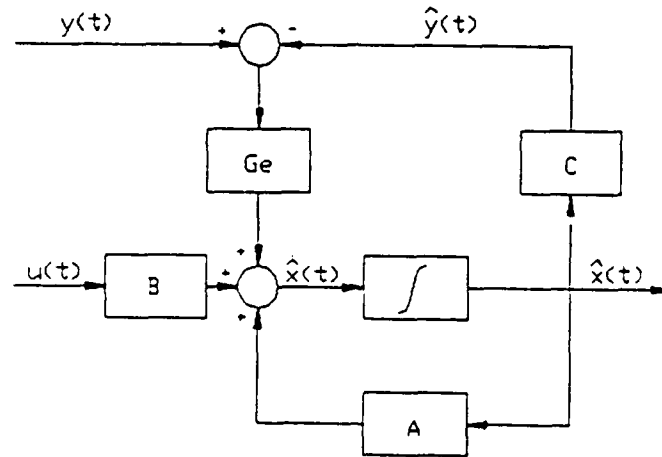


Figure 5.1 Block diagram of a full-order estimator

The state differential equation of the system is

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\bar{\mathbf{u}}(t) \quad (5-1)$$

If the output of the system is $\bar{\mathbf{w}}(L, t)$, we can find a coefficient matrix \mathbf{C}' which is

$$\mathbf{C}' = [0 \ 0 \ \dots \ 1 \ 0 \ 0 \ \dots \ 0]_{1 \times 2N} \quad (5-6)$$

such that

$$y(t) = \bar{w}(L, t) = C'x(t) \quad (5-7)$$

The design goal is to find an estimator taking the $\bar{w}(L, t)$ and $\bar{u}(t)$ as input which can generate output state $\hat{x}(t)$ equal to the real system state $x(t)$ after running the system a finite period of time. Based on this, we can set a state differential equation for the estimator as

$$\dot{\hat{x}}(t) = F\hat{x}(t) + G_e\bar{w}(L, t) + H\bar{u}(t) \quad (5-8)$$

By subtracting Eq. 5-1 from 5-8 and combining with Eq. 5-7, we have the following differential equation for $x(t) - \hat{x}(t)$,

$$\dot{x}(t) - \dot{\hat{x}}(t) = [A - G_e C']x(t) - F\hat{x}(t) + [B - H]\bar{u}(t) \quad (5-9)$$

Because $x(t) = \hat{x}(t)$ after a period of time, from Eq. 5-9 we can obtain

$$\begin{aligned} F &= A - G_e C' \\ H &= B \end{aligned} \quad (5-10)$$

Therefore, Eq. 5-9 can be written as

$$\dot{x}(t) - \dot{\hat{x}}(t) = [A - G_e C'] [x(t) - \hat{x}(t)] \quad (5-11)$$

which is the differential equation of the reconstruction error of the estimator. From Eq. 5-11 we can easily see that the performance of the estimator depends on the matrix $[A - G_e C']$. Then the state differential equation of the estimator from Fig. 5-1 can be written as

$$\begin{aligned} \dot{\hat{x}}(t) &= A\hat{x}(t) + G_e[\bar{w}(L, t) - C'\hat{x}(t)] + B\bar{u}(t) \\ &= [A - G_e C']\hat{x}(t) + G_e\bar{w}(L, t) + B\bar{u}(t) \end{aligned} \quad (5-12a)$$

And the estimated output of the system is $\hat{w}(L, t)$

$$\hat{w}(L, t) = C' \hat{x}(t) \quad (5-12b)$$

The performance of the estimator depends on the eigenvalues of the matrix $[A - G_e C']$ in Eq. 5-12. So far, choosing matrix G_e is still arbitrary. By using the pole placement technique, we can have the required performance for the estimator. The determination of the estimator gain matrix G_e is the main task of the design. Because the design procedure is the same as the one in Section 4.3, we will not repeat it here.

Case Study

We continue the case study with open-loop control system (referring to Eq. 4-19). Let the control force $u(t)$ is equal to zero in this case.

$$\dot{x}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -54.42 & 48.275 & -47.235 & 15.38 \\ -459.15 & 143.41 & -140.42 & 45.73 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \\ .665 \\ 2.021 \end{bmatrix} \bar{F}(t) \quad (5-13a)$$

The output equation is

$$\bar{w}(L, t) = [0 \quad 1 \quad 0 \quad 0] x(t) \quad (5-13b)$$

If the poles of matrix $[A - G_e C']$ are placed at $-3. \pm 0.03i$ and $-2. \pm 0.02i$. The estimator gain matrix calculated is

$$G_e = \begin{bmatrix} 2.748 \\ 8.495 \\ 4.516 \\ 13.603 \end{bmatrix} \quad (5-14)$$

The state differential equation for the estimator is

$$\dot{\hat{\mathbf{x}}}(t) = \begin{bmatrix} 0 & -2.748 & 1 & 0 \\ 0 & -8.495 & 0 & 1 \\ -154.42 & 43.758 & -47.235 & 15.38 \\ -459.15 & 129.81 & -140.42 & 45.73 \end{bmatrix} \hat{\mathbf{x}}(t) + \begin{bmatrix} 2.748 \\ 8.495 \\ 4.516 \\ 13.603 \end{bmatrix} \bar{w}(L, t) \quad (5-15)$$

And the reconstruct output

$$\hat{\bar{w}}(L, t) = [0 \quad 1 \quad 0 \quad 0] \hat{\mathbf{x}}(t) \quad (5-16)$$

The initial-condition response of the real output and estimated output is shown in Fig. 5.2.

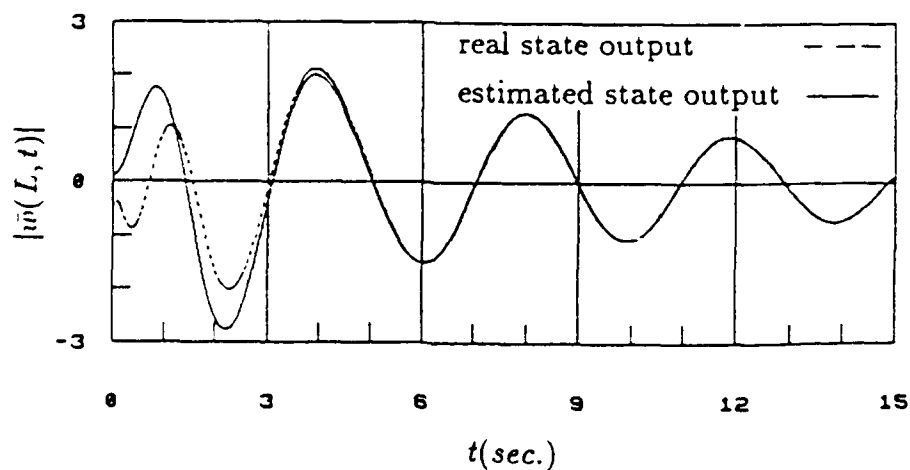


Figure 5.2 Initial-condition response of the real output and estimated output

Output Feedback Control System

In this section we connect the optimal LQR with estimator to have the output feedback control system. This system takes the measured output $y(t)$ for feedback. The overall performance of the output feedback control system depends on both regulator and estimator. From control theory we know that the eigenvalues of the complete system consists of the eigenvalues of regulator and the eigenvalues of estimator separately. This means that we can design regulator and estimator independantly and then put them together.

The state differential equation with optimal LQR is

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\bar{\mathbf{u}}^*(t) + \mathbf{P}(t) \quad (5-17)$$

and the measured output is given by the following equation

$$\bar{w}(L, t) = \mathbf{C}'\mathbf{x}(t) \quad (5-18)$$

The optimal control law calculated in Section 4.4 with estimated state is

$$\bar{\mathbf{u}}^*(t) = -\mathbf{K}^*\hat{\mathbf{x}}(t) \quad (5-19)$$

Also, the estimator designed in preceding section satisfies the following state differential equation

$$\dot{\hat{\mathbf{x}}}(t) = [\mathbf{A} - \mathbf{G}_e\mathbf{C}']\hat{\mathbf{x}}(t) + \mathbf{B}\bar{\mathbf{u}}^*(t) + \mathbf{G}_e\bar{w}(L, t) \quad (5-20)$$

where \mathbf{G}_e is the estimator gain matrix. Fig. 5.3 shows the interconnection of the beam-RMA system, the optimal control law, and estimator.

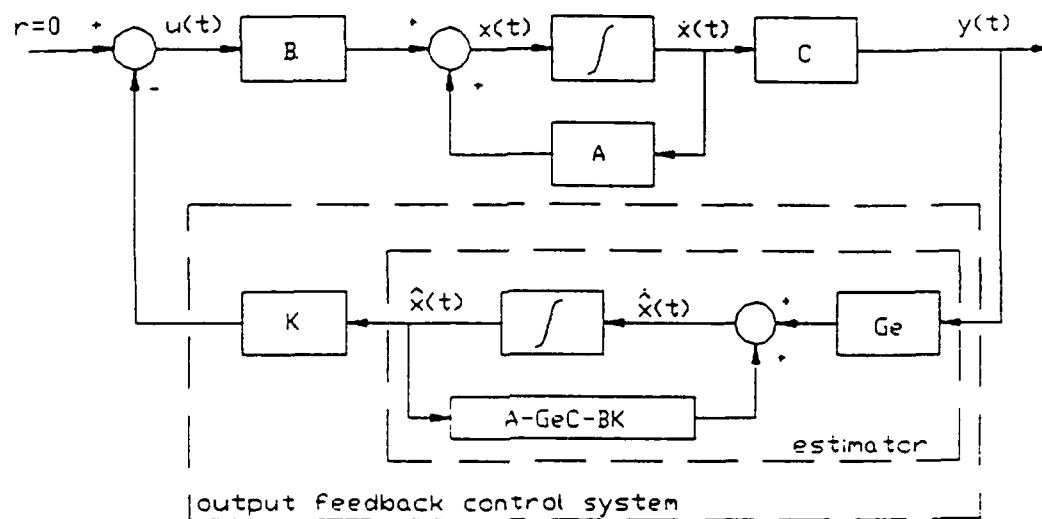


Figure 5.3 The output feedback control system

Combine state differential equations 5-17 and 5-20 with Eqs. 5-18 and 5-19 into one equation. Then the complete system connected with optimal LQG regulator and estimator is described by the state differential equation

$$\begin{bmatrix} \dot{x}(t) \\ \dot{\hat{x}}(t) \end{bmatrix} = \begin{bmatrix} A & -BK^* \\ G_e C' & A - G_e C' - BK^* \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} + \begin{bmatrix} P(t) \\ 0 \end{bmatrix} \quad (5-21)$$

The performance of the complete system depends on the eigenvalues of Eq. 5-21. Based on the linear optimal control theory we know that the poles of the system 5-21 consists of the optimal LQR poles and the estimator poles. Therefore, if both optimal LQR and estimator are well designed, the response of the complete system subjected to the external forces should be satisfactory.

Case Study

From the results of case study in Chapter 4 and preceding section, the state differential equation of complete system can be calculated as

$$\begin{bmatrix} \dot{\mathbf{x}}(t) \\ \dot{\hat{\mathbf{x}}}(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -154.4 & 48.28 & -47.24 & 15.38 & 34.88 & -33.05 & -287.9 & 93.53 \\ -459.2 & 143.4 & -140.4 & 45.73 & 103.8 & -98.32 & -856.5 & 278.2 \\ 0 & 2.75 & 0 & 0 & 0 & -2.75 & 1 & 0 \\ 0 & 8.5 & 0 & 0 & 0 & -8.5 & 0 & 1 \\ 0 & 4.52 & 0 & 0 & -119.5 & 10.7 & -335.1 & 108.9 \\ 0 & 13.6 & 0 & 0 & -355.4 & 31.5 & -996.9 & 324 \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ \hat{\mathbf{x}}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ .665 \\ 2.021 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \bar{F}(t) \quad (5-22)$$

The frequency response and impulse response of the system without and with output feedback control system are shown in Figs. 5.4 and 5.5, respectively.

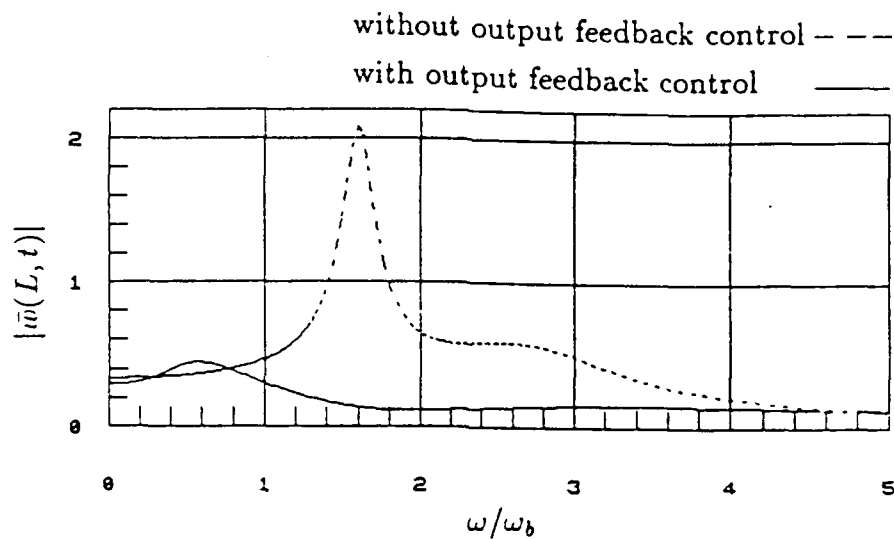


Figure 5.4 Frequency Response of the System without and with Output Feedback Control System

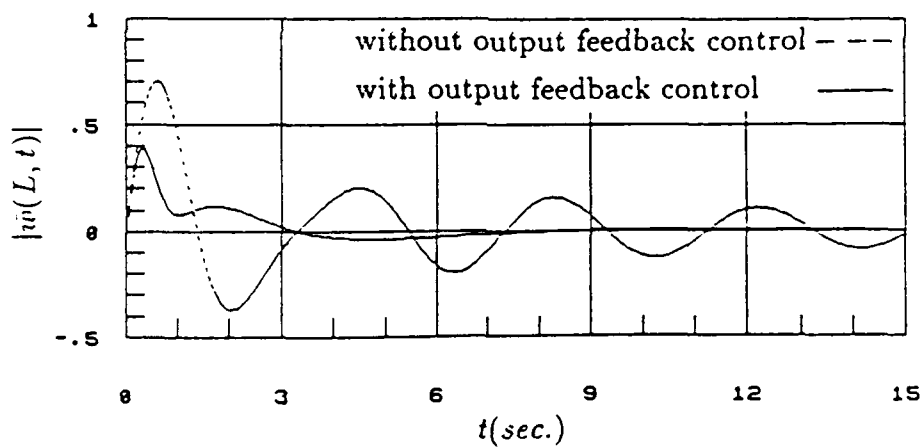


Figure 5.5 Impulse Response of the System without and with Output Feedback Control System

CHAPTER 6

Conclusions

Modal analysis approach has been successfully applied to the cantilever beam-RMA system to solve the boundary-value problem. This approach is more straight-forward in understanding and application than the approach of Green's functions in [2]. Although, the Euler-Bernoulli's beam equation is used, the model derived for the system can be modified to have damping effect in the beam.

Incorporation of RMA to the cantilever beam introduces an additional vibration mode to the system. By adjusting the tuning and damping of RMA, the system response can be changed. RMA used in this way is called a passive vibration absorber (PVA). As we saw in Figs. 3.2 and 3.3 the first mode is split into two modes with little effect on the other modes when the RMA is tuned to the first mode. The same phenomenon happens for the case of tuning to the second mode. Increasing damping in RMA can bring the infinite peak response of resonance down to lower level.

If the system is lightly damped, analytic solution of the undamped system will be a good approximation. Analytic solution of the damped system is not available because of the coupling effect from the damping coefficient matrix in the differential equation. Numerical methods have to be employed to offer a solution. Modal analysis transforms the continuous system to an infinite-degree-of-freedom system which needs to be truncated for numerical calculations. To have more accurate solution, the more modes should be included.

The study of control system design shows that employing RMA to generate control force is a successful approach to reduce the beam vibration. State feedback control system design based on linear optimal control theory improves the performance of the cantilever beam-RMA system. The resonant phenomenon is suppressed to acceptable level. A two-position feedback loop is good enough for reducing the vibration of the first two modes. The optimal feedback gains are suggested to this model. The closed-loop control system also guarantees the ability to compensate the parameter changes in the modeling process. A linear full-order estimator is designed for the control system in the case that the complete state is not available for feedback or can not be accurately measured.

Parameter optimization for the PVA to have best frequency response is in progress. The study in this thesis provides theoretical understanding for experimental implementation which is a task in the future.

Appendix

APPENDIX

Program Listing

The program EIGMAT is written to solve for the eigenvalues and calculate the coefficient matrices of the state differential equation for case study.

```
00100 PROGRAM EIGMAT
00110 REAL FV(25), NF(25,20), EV(20), EVV(20), AN(20),
00115+      C21(20), MASS(20,20),
00120+      DAMP(20,20), STIFF(20,20), MS(20), BB(20),
00125+      RXX(20,20), MSS(25,25),
00130+      W(20), CND(20), MSD(20)
00140 EXTER AL FCN,SHSH,SHCH,CHCH,SHS,SHC,CHS,CHC,SS,
00145+      SC,CC
00150 COMMON /BLOCK1/EV ,/BLOCK2/ALPHA, MU, F, /BLOCK3/
00160+      N3, AN, C21, ZA,MS
00170C PROGRAM IS SET FOR SINGLE TUNING Fa CASE. FOR
00180C MULTI-TUNING CASE, MODIFY LINE 590. ALPHA=POINT
00185C MASS M/MASS OF THE BEAM, MU=MA/M, ZA= C/(2*M*WB),
00190C WB= (EI/DAL4)**0.5, F=(WA/WB), WA=(K/MA)**0.5
00200C MSD IS THE NORMALIZED MODE SHAPE, MS=MSD/CN
00210
00220 ALPHA=0.25
00230 MU    = 1.0
00240 ZA    = 0.5
00250 N3     = 5
00260 N4 = 4*N3 + 1
00270 DF = 1.8
00280 FMAX = 1.8
00290 N1    = INT(FMAX/DF) + 1
00300 X0 = 0.001
00310 DX=0.1
00320 XMAX=15.
00330 N2=INT(XMAX/DX)
00340 XTOL = 0.0001
00350 YTOL = 0.0001
00360 NLIM = 10
00370 DO 10 I= 1, 1
00380     FV(I) = 0
00390 DO 10 J= 1, 20
00400     NF(I,J) = 0
```

```

00410     EV(J) = 0
00420     AN(J) = 0
00430     C21(J) = 0
00440     MS(J) = 0
00450     BB(J) = 0
00460     W(J) = 0
00470     CND(J) = 0
00480     MSD(J) = 0
00490 10 CONTINUE
00500
00510 DO 20 I= 1,20
00520 DO 20 J= 1,20
00530     MASS(I,J)= 0.
00540     DAMP(I,J)= 0.
00550     STIFF(I,J)= 0.
00560     RXX(I,J)= 0.
00570 20 CONTINUE
00580
00590 DO 100 I= N1,N1
00600     K = 1
00610     F = DF * (I-1)
00620     FV(I) = F
00630
00640 DO 100 J = 1, N2
00650
00660     X1 = DX* (J-1) + X0
00670     Y1 = FCN(X1)
00680     X2 = DX* J + X0
00690     Y2 = FCN(X2)
00700 IF ( Y1*Y2 .GT. 0 ) GO TO 100
00710
00720 I1 = 1
00730 CALL MDLNIN (FCN,X1,X2,XR,XTOL,YTOL,NLIM,I1)
00740 NF(I,K) = XR
00750 K=K+1
00760
00770 100 CONTINUE
00780
00790 PRINT 110
00800 110 FORMAT (5(/), 5X, 7HFa      ,14H EIGENVALUES*L)
00810 DO 130 I= 1,N1
00820 PRINT 120, FV(I), ( NF(I,K), K= 1, 6)
00830 120 FORMAT (/ , 7E10.4)
00840 130 CONTINUE
00850
00860C CALCULATING THE COEFFICIENT MATRICES
00870 DO 200 I= 1, N3
00880     EV(I) = NF(N1,I)

```

```

00890  AN(I) = (F**2)/(F**2-EV(I)**4)
00900  C21(I) = -(COSH(EV(I))+COS(EV(I)))/(SINH(EV(I))+SIN
00905+      (EV(I)))
00910  DO 150 J = 1, N4
00920  EVV(J) = (EV(I)/(N4-1)) * (J-1)
00930  MSS(I,J)=(COSH(EVV(J))-COS(EVV(J)))+C21(I)*(SINH
00935+      (EVV(J))-SIN(EVV(J)))
00940  150 CONTINUE
00950  MS(I) = MSS(I,N4)
00960  200 CONTINUE
00970  DO 202 I= 1, N3
00980  DO 202 J= 1, N3
00990  RXX(I,J) = MS(I) * MS(J)
01000  202 CONTINUE
01010
01020  CALL MASSM (SHSH,SHCH,CHCH,SHS,SHC,CHS,CHC,SS,SC,CC,
01025+      MASS )
01030  CALL STIFFM (SHSH,SHCH,CHCH,SHS,SHC,CHS,CHC,SS,SC,CC
01035+      , STIFF )
01040  DO 400 I= 1, N3
01050  CND(I) = (1/MASS(I,I))**0.5
01060  W(I) = STIFF(I,I)/(MASS(I,I))
01070  DO 430 J= 1, N4
01080  MSS(I,J) = CND(I)*MSS(I,J)
01090  430 CONTINUE
01100  MSD(I) = MSS(I,N4)
01110  BB(I) = ( 1 - AN(I))* MSD(I)
01120  400 CONTINUE
01130  CALL DAMPM ( MSD, DAMP )
01140
01150  PRINT 205
01160  205 FORMAT(5(/),45HMAGNITUDES OF NORMALIZED MODE
01170+  SHAPES FROM X=0,42H TO L, DX/L = 0.05 WHICH ALSO
01175+  IS MATRIX G.)
01180  DO 208 I = 1, N3
01190  PRINT 280 , (MSS(I,J),J=1,N4)
01200  208 CONTINUE
01210  PRINT 410
01220  410 FORMAT(3(/),21HNCOEFFICIENT MATRIX Q )
01230C PRINT 280, (MSD(I), I= 1, N3)
01240  PRINT 420
01250  420 FORMAT(3(/),30HDIAGONAL TERMS OF MATRIX OMEGA )
01260  PRINT 280, (W(I), I= 1, N3)
01270C PRINT 210
01280  210 FORMAT(3(/), 11HMASS MATRIX)
01290C DO 220 I= 1, N3
01300C PRINT 280, (MASS(I,J), J=1,N3)
01310C 220 CONTINUE

```

```

01320 PRINT 230
01330 230 FORMAT(5(/),16HDAMPING MATRIX C)
01340 DO 240 I= 1, N3
01350 PRINT 280, (DAMP(I,J), J=1,N3)
01360 240 CONTINUE
01370C PRINT 250
01380C 250 FORMAT(5(/),16HSTIFFNESS MATRIX)
01390C DO 260 I=1,N3
01400C PRINT 280, (STIFF(I,J), J=1,N3)
01410C 260 CONTINUE
01420 280 FORMAT(/, 7(2X,E10.4))
01430C PRINT 290
01440 290 FORMAT (5(/),28HRXX OF THE PERFORMANCE INDEX)
01450C DO 295 I = 1, N3
01460C PRINT 280, (RXX(I,J), J= 1, N3)
01470C 295 CONTINUE
01480 PRINT 300
01490 300 FORMAT(5(/),31HTHE COEFFICIENT MATRIX B )
01500 PRINT 280, ( BB(I), I= 1, N3)
01510 STOP
01520 END
01530
01540 SUBROUTINE MDLNIN (FCN,X1,X2,XR,XTOL,FTOL,NLIM,I)
01550 LOGICAL PRIN
01560 PRIN=.TRUE.
01570 IF (I .NE. 0) PRIN = .FALSE.
01580 F1=FCN(X1)
01590 F2=FCN(X2)
01600 IF (F1*F2 .GT. 0) GO TO 50
01610 FSAVE=F1
01620 DO 20 J= 1,NLIM
01630 XR=X2-F2*(X2-X1)/(F2-F1)
01640 FR=FCN(XR)
01650 XERR=ABS(X1-X2)/2.
01660 IF(.NOT. PRIN) GO TO 5
01670 PRINT 199, J, XR, FR
01680 199 FORMAT(1H,13HAT ITERATION , I4, 5H X = ,E12.5,
01690+ 9H, F(X) = ,E12.5)
01700 5 IF (XERR .LE. XTOL) GO TO 60
01710 IF (ABS(FR) .LT. FTOL) GO TO 70
01720 IF (FR*F1 .LT. 0) GO TO 10
01730 X1 = XR
01740 F1 = FR
01750 IF (FR*FSAVE .GT. 0) F2 = F2/2
01760 FSAVE = FR
01770 GO TO 20
01780 10 X2 = XR
01790 F2 = FR

```

```

01800      IF (FR*FSAVE .GT. 0) F1 = F1/2
01810      FSAVE = FR
01820 20 CONTINUE
01830
01840 I = -1
01850 PRINT 200, NLIM, XR, FR
01860 200 FORMAT (1H0, 26HTOLERANCE NOT MET. AFTER , I4,
01870+ 15H ITERATIONS X = , E12.5, 12H AND F(X) = ,
01875+ E12.5)
01880 RETURN
01890 50 I = -2
01900 PRINT 201
01910 201 FORMAT (1H0, 35HFUNCTION HAS SAME SIGN AT X1
01915+ AND X2)
01920 RETURN
01930
01940 60 I = 1
01950 PRINT 202, J, XR, FR
01960 202 FORMAT (1H0, 19HX TOLERATNCE MET IN , I4,
01965+ 18H ITERATIONS. X = ,
01970+ E12.5, 8H F(X) = , E12.5)
01980 RETURN
01990 70 I = 2
02000 PRINT 203, J, XR, FR
02010 203 FORMAT ( 1H0, 19HF TOLERANCE MET IN , I4,
02015+ 18H ITERATIONS. X = ,
02020+ E12.5, 8H F(X) = , E12.5)
02030 RETURN
02040 END
02050
02060 SUBROUTINE MASSM (SHSH,SHCH,CHCH,SHS,SHC,CHS,CHC,SS,
02065+ SC,CC,M)
02070 REAL A(20), M(20,20), C(20), MS(20)
02080 COMMON /BLOCK2/ AL, MU, F, /BLOCK3/ N, A, C, ZA, MS
02090
02100 DO 100 I = 1, N
02110 DO 100 J= 1, N
02120 M(I,J) = CHCH(I,J) - CHC(I,J) - CHC(J,I) +
02130+ CC(I,J)+C(I)*( SHCH(I,J) - SHC(I,J)
02135+ - CHS(J,I) +SC(I,J) )
02140+ +C(J)*( SHCH(J,I) - SHC(J,I) - CHS(I,J) +
02145+ SC(J,I) )
02150+ +C(I)*C(J)*( SHSH(I,J) - SHS(I,J) -
02155+ SHS(J,I) + SS(I,J) )
02160+ +AL* MS(I)* MS(J)* (1+MU*A(I)*A(J))
02170 100 CONTINUE
02180 RETURN
02190 END

```


MATRIXx Input File

```

***** This is for open-loop system with damping. *****
***** For undamped system, modify matrix A. *****
alpha=0.25; mu=1.0; fa=1.8; zeta=0.5;
A = [ 0 0 1 0 ;
      0 0 0 1 ;
      -154.42 48.275 -47.235 15.38 ;
      -459.15 143.41 -140.42 45.73];
B = [ 0 0 .665 2.021]';
C = [ 0 1 0 0 ]; D=0;
S = [A B;
      C D];
NS=4; NR=[0.01;5;300];
[OME,H] = FREQ(S,NS,NR);
X1 = ABS(H);
TMAX = 15; NPTS = 300; X0 = [0 .1 0 0]';
[T,X3] = IMPULSE(S,NS,TMAX,NPTS);
[T,XI1] = INITIAL(S,NS,X0,TMAX,NPTS);
***** Optimal LQG regulator *****
QC = [ 0 0 .677 2.014]';
RXX = DIAG([0 1 0 0]); RUU = 1E-3;
[EVAL,K] = REGULATOR(A,QC,RXX,RUU)
AK = A - QC*K;
S = [AK B;
      C D];
[OME,H] = FREQ(S,NS,NR);
X2 = ABS(H);
[T,X4] = IMPULSE(S,NS,TMAX,NPTS);
X = [X1 X2];
XIM = [X3 X4];
***** Estimator design *****
POLES = [-3.+ 0.03*JAY, -2.+0.02*JAY];
GE = POLEPLACE(A',C',POLES)
AE = A - GE'*C;
S = [ AE GE';
      C D ];
X0 = [0 -.5 0 0]'; DELTAT = .05;
[T,XI2] = LSIM(S,NS,XI1,DELTAT,X0);
XI = [XI1 XI2];
***** Output feedback control system *****
AO = [ A -QC*K ;
      GE'*C AE-QC*K ];
BO = [ B ; 0 ; 0 ; 0 ; 0 ];
CO = [ C 0 0 0 0 ];
SO = [ AO BO ;
      CO D ];
NS = 8 ;

```

```
[OME,H] = FREQ(SO,NS,NR);  
XO = ABS(H);  
XO = [X1 XO];  
TMAX = 15. ; NPTS = 300 ;  
[T,X5] = IMPULSE(SO,NS,TMAX,NPTS);  
XIO = [X3 X5];
```

References

REFERENCES

- [1] Gerald, C. F., Applied Numerical Analysis. Reading, MA: Addison-Wesley Publishing Co., 1978.
- [2] Kwakernaak, H. and R. Sivan, Linear Optimal Control Systems. New York: Wiley-Interscience, a Division of John Wiley & Sons, Inc., 1972.
- [3] Nicholson, J. W. and L. A. Bergman, Vibration of Combined and/or Constrained Linear dynamical Systems. Report No. T. & A.M. 467. Department of Theoretical and Applied Mechanics, University of Illinois at Urbana Champaign, April 1984.
- [4] Meirovitch, L., Elements of Vibration Analysis. 2nd ed. New York: McGraw-Hill Book Co., 1986.
- [5] Timoshenko, S., D. H. Young, and W. Weaver, Jr., Vibration Problems in Engineering. 4th ed. New York: John Wiley & Sons, Inc., 1974.
- [6] James, M. L., G. M. Smith, J. C. Welford, and P. W. Whaley, Vibration of Mechanical and Structural Systems: with Microcomputer Applications. New York: Harper & Row, Publishers, Inc., 1989.
- [7] Snowdon, J. C., Vibration and Shock in Damped Mechanical Systems. New York: John Wiley & Sons, Inc. 1968.
- [8] Franklin, G. F., J. D. Powell, and A. Emami-Naeini, Feedback Control of Dynamic Systems. Reading, MA: Addison-Wesley Publishing Co., 1987.
- [9] Ogata, K., Modern Control Engineering. Englewood Cliffs, NJ: Prentice-Hall, Inc. 1970.